TESTS OF SIGNIFICANCE IN FACTOR ANALYSIS

By M. S. BARTLETT
Department of Mathematics, University of Manchester

I. *Description and Illustration of a Test of Significance for Analysis into Principal Components.* II. *Remarks on Lawley's Maximum Likelihood Method.* III. *Further Theoretical Discussion.*

I. DESCRIPTION AND ILLUSTRATION OF A TEST OF SIGNIFICANCE FOR ANALYSIS INTO PRINCIPAL COMPONENTS

Introductory Remarks. In a previous paper (1) in this *Journal* I distinguished between two types of statistical analysis, external canonical or factor analysis between two groups of variables and internal factor analysis of a single group. By the latter was meant the particular statistical analysis of the internal correlation structure known more familiarly as an ' analysis into principal components.' It was shown that the appropriate tests of significance of the ' factors ' emerging from the statistical analysis were quite distinct in the two cases ; but while a fairly satisfactory test was available for the external case, no analogous test was given in the internal case, only a method of assessing the significance of the smallest root but one against the smallest. It is the main purpose of the present paper to remedy this deficiency, and describe a more general test analogous to the test developed for the external case.

It should be noticed that an analysis into principal components is distinct from an analysis corresponding to some postulated psychological pattern, and has a different interpretation. The most systematic attempt to cover the latter problem is perhaps that by Lawley (2, 3, 4), using the method of maximum likelihood. A critical discussion of this method is given in Part II, where it is shown further that the associated χ^2 test proposed by Lawley in 1940 has a partial connexion with the χ^2 technique proposed in I. For convenience the theoretical justification of the formulæ and methods described in I (and, to some extent, in II) is deferred to Part III, which may be omitted by the non-mathematical reader.

The present paper does not entirely clear up the problems which it raises for either form of analysis ; but the procedures discussed seem sufficiently justified to be worth reviewing at this stage, even although further theoretical investigation is still needed.

The χ^2 Approximation. Let us denote by λ_i ($i = 1, 2, \ldots, p$) the p roots, in descending order of magnitude, of the determinantal equation

$$| R - \lambda | = 0 , \qquad (1)$$

where R is a correlation matrix of p variables. It should be noted that for the determinant $| R |$ we have the relation $| R | = \lambda_1 \lambda_2 \ldots \lambda_p$. Let the total number of degrees of freedom available from the original observations be n ; we shall usually have $n = r - 1$, when deviations from the general mean are employed for each variable, where r is the total number of observations for each variable, i.e., in the present context r is the total number of persons tested.[1] The statistical significance

[1] Note that the number of degrees of freedom s of Table I of the previous paper was defined as $r - 2$. This is the usual convention for the test of significance of a single correlation, but the degrees of freedom $n = r - 1$ for each variable is more appropriate here.

of the entire correlation structure may be tested if necessary by calculating the quantity [1]

$$\chi^2 = -\{n - \tfrac{1}{6}(2p + 5)\} \log_e | R | \ , \qquad (2)$$

with $\frac{1}{2}p(p-1)$ degrees of freedom. However, the usual situation will be that after the removal of the largest roots we require to test the significance of those left. The χ^2 test may still, as in an external analysis, be used for this purpose, provided that the roots removed correspond to well-determined (i.e., highly significant) factors. We now take

$$\chi^2 = -\{n - \tfrac{1}{6}(2p + 5) - \tfrac{2}{3}k\} \log_e R_{p-k} \, , \qquad (3)$$

with $\frac{1}{2}(p-k)(p-k-1)$ degrees of freedom, after k roots $\lambda_1 \ \lambda_2 \ldots \lambda_k$ have been determined, where

$$R_{p-k} = | R | \Big/ \left\{ \lambda_1 \lambda_2 \ldots \lambda_k \left[\frac{p - \lambda_1 - \lambda_2 \ldots - \lambda_k}{p-k} \right]^{p-k} \right\} . \qquad (4)$$

It will be noticed that the test is formulated to indicate the significance of the residual roots ; however, it is often convenient to present a complete χ^2 table, provided this is properly interpreted. Since the multiplying factor in (3) changes with k, the complete table is not strictly additive unless a constant factor is substituted ; we might employ the smallest, corresponding to $k = p - 2$: the factor is then $n - p + \frac{1}{2}$. But if n is not large, the full advantage of the approximation is retained if the proper factor is used for each individual test.

Numerical Illustrations. Two examples will be given to illustrate this test. For the first I have taken the data of Truman Kelly's quoted in Hotelling's original discussion (5) of an analysis into principal components, consisting of four tests (reading speed, reading power, arithmetic speed, and arithmetic power) carried out on 140 children. For the matrix R we have [2]

$$\begin{pmatrix} 1 & 0{\cdot}698 & 0{\cdot}264 & 0{\cdot}081 \\ 0{\cdot}698 & 1 & -0{\cdot}061 & 0{\cdot}092 \\ 0{\cdot}264 & -0{\cdot}061 & 1 & 0{\cdot}594 \\ 0{\cdot}081 & 0{\cdot}092 & 0{\cdot}594 & 1 \end{pmatrix}$$

The variances taken up by the four roots were

λ_1	=	1·846		
λ_2	=	1·465	$\lvert R \rvert$	= $\lambda_1 \lambda_2 \lambda_3 \lambda_4$
λ_3	=	0·521		= 0·2353
λ_4	=	0·167		
Total	=	3·999		

For the quantities R_{p-k} in (3) we find

$$R_2 = 0{\cdot}521 \times 0{\cdot}167 \times (2/0{\cdot}689)^2 = 0{\cdot}7330$$
$$R_3 = (0{\cdot}2353/1{\cdot}846) \times (3/2{\cdot}154)^3 = 0{\cdot}3444$$
$$R_4 = | R | = 0{\cdot}2353 \ .$$

In practice when the smallest roots have not been evaluated $| R |$ must of course be evaluated directly. Using the common factor $n - p + \frac{1}{2} = 139 - 4 + \frac{1}{2} = 135{\cdot}5$, we obtain the additive χ^2 analysis :

[1] As noted in Part III, the criterion $| R |$ was considered by Wilks in 1932 ; the χ^2 approximation proposed in equation (2) is closely connected with the χ^2 approximation I have put forward for external analysis (see also Box (11) p. 344).

[2] In order to make use of Hotelling's analysis for this first example these correlations, corrected for attenuation, have been used, though of course this correction will affect test of significance to some (unknown) extent which is here ignored.

	D.F.	χ^2
λ_1	3	$-135{\cdot}5$ ($\log_e 0{\cdot}2353 - \log_e 0{\cdot}3444$) = 51·6
λ_2	2	$-135{\cdot}5$ ($\log_e 0{\cdot}3444 - \log_e 0{\cdot}7330$) = 102·3
λ_3	1	$-135{\cdot}5$ ($\log_e 0{\cdot}7330$) = 42·1
λ_4	—	
Total ..	6	$-135{\cdot}5$ ($\log_e 0{\cdot}2353$) = 196·0

The best approximation for the total χ^2 would actually be $-\{n - \frac{1}{6}(2p + 5)\} \log_e |R| = -136{\cdot}833$ ($\log_e 0{\cdot}2353$) = 198·0, but of course all the components are so highly significant that this difference is unimportant. It may be wondered why the second largest component gives an apparently more significant χ^2 than the first; but it should be remembered that logically the significance of at least one root is to be judged from the total χ^2, not the component corresponding to λ_1. The apparent significance of this component according to the above table represents the magnitude of the root compared with the variation still remaining, and the latter contains quite a large second component. On the other hand, the apparent significance of the latter according to the analysis of the table is judged in relation to the residual variation after this component has in turn been removed, and this residual variation is comparatively small.

As the second example I have taken the hypothetical data discussed by Burt (6), for which the number of persons is only 6. It is indicated in the theoretical discussion in III that we may expect the test presented here to work satisfactorily, at least for the total χ^2, for as low as about 10 degrees of freedom.[1] This means that the test must be used with considerable caution in the present example, which is unrealistic in its very small number of persons. In view of the small value of n, the χ^2 table for this example will be presented not in additive form, but in the form in which it would arise as each set of residual factors is tested in turn for significance, the best multiplying factor given in (3) being used in each case.

From Table XIII of Burt's paper we have

$\lambda_1 = 3{\cdot}1021$
$\lambda_2 = 0{\cdot}7328$
$\lambda_3 = 0{\cdot}1593$
$\lambda_4 = 0{\cdot}0058$
Total $= 4{\cdot}0000$

$|R| = \lambda_1 \lambda_2 \lambda_3 \lambda_4 = 0{\cdot}002101$

Hence

$$R_2 = 0{\cdot}1593 \times 0{\cdot}0058 \times (2/0{\cdot}1651)^2 = 0{\cdot}1356$$
$$R_3 = (0{\cdot}002101/3{\cdot}1021) \times (3/0{\cdot}8979)^3 = 0{\cdot}02526$$
$$R_4 = |R| = 0{\cdot}002101 .$$

The χ^2 test for the residual roots is thus :

	D.F.	χ^2	$P = 0{\cdot}05$ level
$\lambda_1 \lambda_2 \lambda_3$	3+2+1	$-2{\cdot}833$ ($\log_e 0{\cdot}002101$) = 17·47	12·59
$\lambda_2 \lambda_3$	2+1	$-2{\cdot}167$ ($\log_e 0{\cdot}02526$) = 7·97	7·82
λ_3	1	$-1{\cdot}5$ ($\log_e 0{\cdot}1356$) = 3·00	3·84

[1] By these degrees of freedom is meant of course the number depending on the number of persons, *not* the degrees of freedom for χ^2, depending only on the number of tests.

Even with such a small sample we are thus able to conclude that the first two factors are significant ; the third, however, does not reach significance (cf. the conclusion in the final paragraph of my previous paper (1)).

Further Comments. It need hardly be stressed that I am concerned here with *statistical* significance. It does not follow that all the factors which reach statistical significance in a large sample necessarily remove a very large fraction of the variance ; and hence some of them may be comparatively unimportant in practice. Again, even if they are numerically important, this has no necessary implications of psychological or other reality of the factors. Merely the correlation structure of the variables is being investigated in its relation to variance : for this reason no significance can ever be attached to the last root, for it would be equivalent to asking for the correlation structure of a single variable.

If of course the basic equation (1) is replaced by another, then the above tests become inappropriate ; for example, if we ask for the significance of the factors of an equation like

$$| R - \lambda e_{ii} | = 0 , \qquad (5)$$

where e_{ii} denotes error variances (assumed known), then, as I showed in (1), the analysis is a special case of an external factor analysis, and other tests become relevant. Or if in place of the empirical analysis of equation (1) we substitute a factor analysis of more orthodox psychological type, then, as noted in the introductory remarks, other tests again will be required. Some discussion is given in Part II of a factor analysis of standard type, which in vector and matrix notation is represented by the equation

$$T = M_0 F_0 + M_1 F_1 , \qquad (6)$$

where T represents the vector of standardized tests, F_0 the set of (standardized) general or group factors, F_1 the (standardized) specifics, and (M_0 , M_1) the factor 'loadings.' Since the difference between the factors of equation (6) and those corresponding to the analysis (1) is not always realized, it may be advisable to illustrate this difference by a simple example. Consider three tests of which the first two are correlated with coefficient ρ (positive), and the third is entirely independent. On the basis of equation (6) this might be interpreted as indicating one 'general' factor common to the first two tests, but with zero loading for the third, the remaining factors being specific to each test. On the basis of an analysis into principal components, however, it is easily seen that the roots are $1 + \rho$, 1 and $1 - \rho$, and it will be noticed that in order of contribution to total variance it is the second root which corresponds to the uncorrelated test and, for any reasonable size of sample, this will be significant compared with the last. Thus once significant correlation is present it may automatically cause all the roots to be distinct in a principal components analysis, which should not be assumed to lead necessarily to the *simplest* factor-pattern.

II. REMARKS ON LAWLEY'S MAXIMUM LIKELIHOOD METHOD

The Maximum Likelihood Method in Factor Analysis. To determine the structural factor equation of the type represented by equation (6), Lawley has investigated two methods based on the principle of maximum likelihood. The first, which he has called 'Method I,' is based on the variance-covariance sampling distribution of the test scores ; the second, 'Method II,' is based more directly on the individual test scores (see his paper (3)). At first sight the second method would

appear the more fundamental, but it is known to lead to difficulties. In a recent paper (7) M. G. Kendall raised doubts on the suitability of the maximum likelihood method in the present context ; and Dr. Lawley has informed me that he had already entertained similar doubts, at least as regards ' Method II.' In the discussion following Kendall's paper, I pointed out that the factor equation (6) is equivalent to another set of structural equations (familiar in econometrics) for which it is definitely known that the standard maximum likelihood method breaks down.

The question remains of the validity and status of Lawley's ' Method I,' and a detailed re-examination and formulation of the most appropriate statistical procedure for dealing with the structural equation (6) seems still required. Some provisional comments on Lawley's first method may, however, be useful.

Lawley's ' Method I,' and its Associated χ^2 *Test.* I was at first sceptical of the value of this method because of its obvious breakdown in elementary cases. For example, it does not work for the case of two tests and one general factor, for we should then have three observed quantities, the two variances and the covariance, and four unknown loading coefficients (or equivalently, when the tests have been standardized, one observed correlation, and two unknown loading coefficients for the general factor) ; in other fields this case often arises in the guise of fitting a straight line to the plot of two variables both of which are subject to error (see, for example, Bartlett (8)).

In psychology the number of tests is usually large enough for this difficulty not to arise. The validity of the maximum likelihood method applied to the variance-covariance distribution then depends (i) on our deciding to consider only the information contained in the variance-covariance (or equivalent correlation) matrix, and (ii) on Lawley's procedure, given (i), providing efficient estimates. Since the general validity of the maximum likelihood approach can no longer be assumed, (ii) is no longer automatic, and appears to require demonstration. One possible argument is to note that enough factors and unknown coefficients can always be introduced to take up all the available degrees of freedom and yet avoid redundancy ; representation of the correlation matrix by still fewer parameters merely implies further linear restrictions among the unknown coefficients, and this should hardly affect the question of efficiency when the likelihood method is used.

But it might be noticed that the method depends on the correlation matrix being compatible with the postulated factor-pattern. For example, with only three tests it is not possible to have a non-redundant factor equation (6) except in the case of only one general factor, which should take up all the degrees of freedom. Yet the correlation matrix

$$\begin{pmatrix} 1 & 0{\cdot}60 & -0{\cdot}28 \\ 0{\cdot}60 & 1 & 0{\cdot}60 \\ -0{\cdot}28 & 0{\cdot}60 & 1 \end{pmatrix}$$

is only compatible with a minimum of two general factors. This would have to be revealed by equations impossible to satisfy with real coefficients, and not by Lawley's associated χ^2 test. If the above matrix arose in a sample, it is not clear how we should test the goodness of fit of one general factor. If we ignore such difficulties, Lawley's method appears to provide a satisfactorily objective procedure for dealing with equation (6).

It has been stressed that this factor analysis is quite distinct from that discussed in Part I of this paper ; but it is an attractive link between the two analyses that the total χ^2 corresponding to the significance of the unreduced correlation matrix is necessarily the same, and only because of the difference between the factors extracted in the two analyses does the analysis of the total χ^2 into its respective components differ.

It follows (see the last section of Part III of this paper) that the form of Lawley's test which should be more exact for moderate-sized samples is the determinantal form

$$\chi^2 \sim -n' \log_e | R | / | \hat{R} | ,$$

where $| \hat{R} |$ denotes the estimated correlation matrix when k general factors (and p specifics) have been fitted, and the multiplying coefficient n' be taken, not to be n, but the coefficient $n - \frac{1}{6}(2p + 5)$ used in (2), or perhaps even better, the coefficient $n - \frac{1}{6}(2p + 5) - \frac{2}{3}k$ used in (3). The number of degrees of freedom of χ^2 is $\frac{1}{2}(p - k)(p - k - 1) - k$ (subject, of course, to the difficulties raised above).

III. FURTHER THEORETICAL DISCUSSION

Equivalent Analyses of Correlation Structure. Any resemblance of the test described in I above to the test for external analysis described in the earlier paper is more than superficial. Let us consider the logical relations among a group of variables, say, four for definiteness. The possible correlations may be considered in more than one way. Thus we may consider the correlation between variables x_1 and x_2, then the relation between x_3 and the group x_1, x_2, then between x_4 and the group x_1, x_2, x_3. As another alternative, we might consider the external relation between the two groups x_1, x_2 and x_3, x_4 ; then the further internal relations between x_1 and x_2 in the first group, and between x_3 and x_4 in the second. The internal relations in a single group can thus be regarded as a number of ' external ' relations between various subgroups. Hence a test of the internal relations can be built up from the tests of the equivalent external relations between sub-groups. If we carry out this theoretical programme according, say, to the first of the alternative analyses mentioned above, where the relation of each further variable is considered successively, we arrive from the theory of the χ^2 test for external analysis (9) at the successive χ^2 quantities :

$$\begin{aligned} &-\{n - \tfrac{1}{2}(3)\} \log_e (1 - r_{12}^2) , && (1 \text{ d.f.}) , \\ &-\{n - \tfrac{1}{2}(4)\} \log_e (1 - R_{3\,,\,12}^2) , && (2 \text{ d.f.}) , \\ &\quad\vdots \\ &-\{n - \tfrac{1}{2}(p + 1)\} \log_e (1 - R_{p\,,\,12\ldots p-1}^2) , && (p - 1 \text{ d.f.}) , \end{aligned}$$

where r_{12} is the observed correlation between x_1 and x_2, $R_{p\,,\,12\ldots p-1}$ the multiple correlation of x_p with x_1 , x_2 , . . . , x_{p-1}. In order to make any test arrived at invariant to the route of arrival, it is obviously convenient to choose a mean multiplying factor, which (weighting with the degrees of freedom) becomes $n - \frac{1}{2}f$, where

$$\tfrac{1}{2}p(p - 1)f = (1 \times 3) + (2 \times 4) + (3 \times 5) + \ldots (p - 1)(p + 1) ,$$

or

$$f = \tfrac{1}{3}(2p + 5) .$$

It is well known that

$$(1 - r_{12}^2)(1 - R_{3\,,\,12}^2) \ldots (1 - R_{p\,,\,12\ldots p-1}^2) = | R | ,$$

whence we obtain as an approximate χ^2, with $1 + 2 + \ldots + (p - 1) = \frac{1}{2}p(p - 1)$ degrees of freedom, the expression proposed in equation (2).

Direct Derivation of the χ^2 Approximation. The above method of approach was the one I used first to arrive at (2) ; but it is satisfactory that the test finally arrived at depends on the correlation determinant $| R |$, a criterion considered many years ago by Wilks (10). It is useful to check directly the appropriateness of the χ^2 approximation by writing down the moment-generating function of $-n \log_e | R |$, by suitably adapting Wilks' formula

for the sth order moment of $|R|$ (cf. the derivation in (9) of the χ^2 test for an external analysis). Alternatively the moment-generating function could be written down from the simultaneous distribution of the correlations R, or from the independence of the components listed above. We obtain

$$M(t) = \frac{\Gamma^{p-1}(\frac{1}{2}n) \prod_{i=1}^{p-1} \Gamma\{\frac{1}{2}(n-i) - nt\}}{\Gamma^{p-1}(\frac{1}{2}n - nt) \prod_{i=1}^{p-1} \Gamma\{\frac{1}{2}(n-i)\}}, \qquad (7)$$

and it is a matter of straightforward algebra coupled with Stirling's approximation formula for a Γ-function to show that

$$M(t) = \sum_{i=1}^{p-1} \left\{ t\left[i + \frac{i^2+2i}{2n} + \ldots\right] + t^2\left[i + \frac{i^2+2i}{n} + \ldots\right] + \frac{4}{3}t^3\left[i + \frac{3(i^2+2i)}{2n} + \ldots\right] + \ldots \right\}$$

$$= \tfrac{1}{2}p(p-1)\,t\left[1 - \frac{1}{6n}(2p+5)\right] + \tfrac{1}{2}p(p-1)\,t^2\left[1 - \frac{1}{6n}(2p+5)\right]^2$$

$$+ \tfrac{1}{2}p(p-1)\,\frac{8t^3}{3!}\left[1 - \frac{1}{6n}(2p+5)\right]^3 + \ldots \qquad (8)$$

to $0\left(t^3, \frac{1}{n}\right)$. Hence the improved approximation is

$$\chi^2 \sim -\{n - \tfrac{1}{6}(2p+5)\} \log_e |R| ,$$

confirming the mean multiplying factor arrived at by the previous argument, and from (8) having $\frac{1}{2}p(p-1)$ degrees of freedom.[1]

Closeness of the χ^2 Approximation. The order of closeness of this approximation will clearly be comparable to that for an external analysis, in view of their relationship, and will consequently be expected to work quite well down to a sample with about 10 degrees of freedom (9). Let us, for example, in the case of only two variables, where merely the test of a single correlation is involved, check the closeness of the approximation. We find for χ^2, corresponding to the correct critical $P = 0{\cdot}05$ or $0{\cdot}01$ value of r_{12}, the values :

	$P = 0{\cdot}05$	$P = 0{\cdot}01$
$n = 10$	3·84	6·60
$n = 20$	3·84	6·63
$n = \infty$	3·841	6·635

Effect of Eliminating the Larger Roots. The approximate effect of eliminating the larger roots in the factor analysis has next to be considered. To some extent it is possible to give an argument analogous to that used for an external analysis, when the roots eliminated are well determined. In this case the orthogonal directions corresponding to these roots in the geometrical representation of the variables in the n-dimensional sample space will be well determined, and we may consider the analysis of the remaining $p - k$ roots in a space of $n - k$ dimensions (if k roots have been eliminated).

The adjustment of the multiplying factor gives

$$(n-k) - \tfrac{1}{6}\{2(p-k) + 5\} = n - \tfrac{1}{6}(2p+5) - \tfrac{2}{3}k ,$$

as given in (3). The expression for R_{p-k} given in (3) is merely a convenient method of obtaining the expression in the remaining roots analogous to the determinant $|R| = \lambda_1 \lambda_2 \ldots \lambda_p$, when we are given $|R|$ and $\lambda_1 \lambda_2 \ldots \lambda_k$. It should be remembered that significance in an internal analysis can only be on the basis of the relative values of the roots, which must after each elimination be in effect re-scaled to unit mean variance. This explains the factor $[(p-k)/(\lambda_{k+1} + \ldots \lambda_p)]^{p-k}$ in R_{p-k}, which may thus be written

$$\frac{|R|}{\lambda_1 \lambda_2 \ldots \lambda_k} \left\{\frac{p-k}{p - \lambda_1 - \lambda_2 \ldots - \lambda_k}\right\}^{p-k} .$$

[1] An independent and prior mention of this approximation is included in the recent comprehensive paper on such approximations by Box (11).

It must be admitted that the above justification of the proposed test for the residual roots is not as complete as could be wished. The problem would be slightly simpler if we were alternatively considering the analysis of test scores known to have true equal variances, but standardized to unity only for the *mean* variance. In this case the above form of the test for the residual roots can be given further justification, but of course the total χ^2 now has $p - 1$ more degrees of freedom, these representing the test of homogeneity of the p test variances about their unit mean variance. Correspondingly, the total χ^2 for the last $p - k$ factors would have (approximately) $\frac{1}{2}(p - k - 1)(p - k + 2)$ degrees of freedom instead of $\frac{1}{2}(p - k)(p - k - 1)$. One point that remains in some doubt pending a more detailed examination is whether the reduction in degrees of freedom that ensues from the individual standardization of the tests is automatically felt in the residual factor components (an assumption implicit in the proposed test), or is mainly absorbed by the larger roots.

The Alternative Forms of Lawley's χ^2 Test. In his 1940 paper (2), Lawley gave two alternative forms for the large-sample χ^2 test he developed for use with his maximum likelihood method (Method I). One was an arithmetically convenient sum of squares, but it is worth pointing out that the equivalent determinant formula arises more exactly than he perhaps emphasized, and should therefore be the more accurate one for moderate-sized samples.

To demonstrate this, it is convenient to define a 'homogeneous' likelihood function from the variance-covariance sampling distribution by its logarithm L, where

$$-2L + \text{constant} = n \log |C| - n \log |A| + n \operatorname{trace}(C^{-1}A) . \quad (9)$$

In this expression, where C denotes the true variance-covariance matrix and A the corresponding sample matrix, $-2L$ tends to χ^2 with $\frac{1}{2}p(p + 1)$ degrees of freedom as n increases. To determine the constant in (9) we substitute the estimates A for C, giving

$$\text{constant} = n \operatorname{trace}(A^{-1}A) = np . \quad (10)$$

If we alternatively substitute estimates of C from (efficient) factor estimates, with q independent unknown coefficients, we obtain

$$-2\hat{L} + \text{constant} = n \log |\hat{C}| - n \log |A| + n \operatorname{trace}(\hat{C}^{-1}A) , \quad (11)$$

with $\frac{1}{2}p(p + 1) - q$ degrees of freedom.

From (10) and (11),

$$\begin{aligned} -2\hat{L} &= -n \log |A| / |\hat{C}| + n \operatorname{trace}(\hat{C}^{-1}A) - np \\ &= -n \log |A| / |\hat{C}| + n \operatorname{trace}(\hat{C}^{-1}A - \hat{C}^{-1}\hat{C}) \\ &= -n \log |A| / |\hat{C}| , \end{aligned} \quad (12)$$

the last term vanishing when $\hat{C}$ corresponds to the maximum likelihood estimates (see Lawley (2)).

If no general factors are fitted, the estimates $\hat{C}$ consist simply of the specific factor variances coinciding with the test variances, and (12) reduces to $-n \log |R|$. It is then, as we should expect, identical with the total χ^2 proposed in the principal components analysis, and the modified multiplying coefficient $n - \frac{1}{6}(2p + 5)$ is consequently more accurate. When k general factors have also been fitted, (12) may be expressed in terms of the ratio of the observed and estimated *correlation* determinants. The more accurate multiplying coefficient $n - \frac{1}{6}(2p + 5)$ should still be more appropriate than the crude coefficient n, and the further modified coefficient $n - \frac{1}{6}(2p + 5) - \frac{2}{3}k$ would have some justification here also, though this last refinement is of course not very firmly established. The number of degrees of freedom comes out at $\frac{1}{2}(p - k)(p - k - 1) - k$, the number lost per component being one more than in the principal components analysis.

Finally, it might be noticed that an alternative analysis of the total χ^2 along the lines given in the first section of Part III would sometimes be of practical interest, and since, as was there pointed out, it would then rest on the known χ^2 approximation for *external* analyses, such an analysis would have a somewhat more precise justification than it has yet been found possible to give either for the principal components analysis or for the maximum likelihood factor analysis.

REFERENCES

1. Bartlett, M. S. (1948). ' Internal and external factor analysis.' *Brit. J. Psych., Stat. Sect.*, I, 73.
2. Lawley, D. N. (1940). ' The estimation of factor loadings by the method of maximum likelihood.' *Proc. Roy. Soc. Edin.*, IX, 64.
3. Lawley, D. N. (1942). ' Further investigations in factor estimation.' *Proc. Roy. Soc. Edin.*, LXI, 176.
4. Lawley, D. N. (1949). ' Problems in factor analysis.' *Proc. Roy. Soc. Edin.*, LXII, 1.
5. Hotelling, H. (1933). ' Analysis of a complex of statistical variables into principal components.' *J. Educ. Psych.*, XXIV, 417 and 498.
6. Burt, C. (1947). ' Factor analysis and analysis of variance.' *Brit. J. Psych., Stat. Sect.*, I, 3.
7. Kendall, M. G. (1950). ' Factor analysis as a statistical technique.' Symposium on factor analysis to be published in *J. Roy. Statist. Soc.* (Series B).
8. Bartlett, M. S. (1949). ' Fitting a straight line when both variables are subject to error.' *Biometrics*, 5, 207.
9. Bartlett, M. S. (1938). ' Further aspects of the theory of multiple regression.' *Proc. Camb. Phil. Soc.*, XXXIV, 33.
10. Wilks, S. S. (1932). ' Certain generalizations in analysis of variance.' *Biometrika*, XXIV, 471.
11. Box, G. E. P. (1949). ' A general distribution theory for a class of likelihood criteria.' *Biometrika*, XXXVI, 317.

A METHOD OF STANDARDIZING GROUP-TESTS

By D. N. LAWLEY
University of Edinburgh

Introduction.—A method has been given by Thomson (1) for standardizing group-tests, by means of which the scores obtained by a complete year-group of children in a test can be converted into 'intelligence quotients' or I.Q.s. These I.Q.s are not true quotients; but are similar to Binet I.Q.s in that they make allowance for difference of age, and are distributed normally about a mean of 100. The standard deviation of group-test I.Q.s has for convenience been fixed at 15, to agree very roughly with that of Binet I.Q.s.

The method of standardization depends essentially on finding for each month of age the 5, 16, 50, 84, and 95 percentiles of the distribution of 'raw scores.' Straight lines are then fitted to obtain for each month the estimated value of each of the percentiles. This means that for any given age within the range considered we can estimate the raw scores corresponding to I.Q.s of (approximately) 75, 85, 100, 115, and 125. Scores corresponding to other I.Q.s may then be estimated by interpolation or extrapolation.

The conversion tables so constructed suffer from the disadvantage that raw scores, and not I.Q.s, are given in the body of the table. In general the I.Q. corresponding to a given age and score can only be found by inverse interpolation. It was for this reason that the author developed a somewhat different method of standardization. This has for some years been used to standardize Moray House group-tests, and has been found to give results very close to the former method, with considerably less calculation.

Description of Method.—In developing the method the main idea has been to make the calculations as simple as possible, without too great loss of accuracy. The efficiency, in a statistical sense, could doubtless be increased, but not without adding materially to the time required. Errors of estimation arising from the construction of the conversion table are in most cases negligible compared with the errors of measurement of any mental test, however good.

We shall suppose that data are available for a complete year-group of children in some area, and that the total number of cases is large, say, at least a thousand. It will be assumed that the age a of a child is measured in units of completed months. We shall thus have twelve score distributions, one for each value of a. We may further assume that there are approximately the same number in each month-group.

Let us denote by x the 'raw score' of any child; and let us suppose that the age a is measured from the mean of the year-group. It will then be presumed that the raw score x depends only on the age a and on the I.Q. y of the child, which is, however, unknown. Thus we may write $x = \varphi(y, a)$, a relation which may alternatively be expressed in the form $y = F(x, a)$.

Now it is clear that there is no logical necessity for $F(x, a)$ to be a simple function of x and a. In practice, however, we may usually assume, to a sufficiently good approximation, that $y = f(x) - \beta a$, where $f(x)$ is some function of x only, while β is a positive constant not dependent on either x or a. Furthermore, the I.Q. y is to be chosen in such a way that it will be distributed normally with a mean of 100 and a standard deviation of 15. Since

$$f(x) = y + \beta a \qquad (1)$$

it follows that, for a given value of a, $f(x)$ must be distributed normally with a mean of $(100 + \beta a)$ and a standard deviation of 15.

When a is not fixed, the distribution of $f(x)$ will not be normal, since that of a is not normal but rectangular. We may nevertheless treat $f(x)$ as being very nearly normally

distributed, since the term βa is small ; and as y and a are distributed independently, the total variance σ_f^2 of $f(x)$ is given by $\sigma_f^2 = \sigma_y^2 + \beta^2\sigma_a^2$, where σ_y^2 and σ_a^2 are the variances of y and a respectively. Putting $\sigma_y^2 = 225$ and $\sigma_a^2 = 12$ (since the range of a is 12), we thus have

$$\sigma_f^2 = 225 + 12\beta^2 . \qquad (2)$$

It is first of all necessary to estimate the constant β. To do this we take a given value of x, say, x_0, and calculate for each value of a, i.e., for each month-group, the proportion p of children with scores less than x_0. For greatest accuracy it is desirable to choose the value of x_0 to be near the median of the total score distribution. We then find for each value of p a corresponding value of z, chosen in such a way that p is the probability of a normal variate with zero mean and unit variance lying within the range $-\infty$ to z. For this purpose Karl Pearson's (2) Table I will be found useful.

For each value of a the quantity $100+15z$ will now provide an estimate of the I.Q. y corresponding to a score of x_0. In order to obtain the relationship between z and a for constant x we fit a straight line to the twelve values of z by the method of least squares (with equal weights). The slope of this line when multiplied by -15 gives us an estimate of β.

We now consider the total distribution of raw scores found by pooling those for the twelve month-groups. The scores will usually be grouped in intervals of 5 or 10 marks, and for the score corresponding to the top end of each interval we find the proportion P of children in the whole year-group obtaining less than this. We may then, if we choose, plot P against x and draw a free-hand curve through the resulting points. In actual practice it is easier to employ some simple, non-graphical method of smoothing the values of P.

For a set of equally spaced values of x we thus obtain corresponding proportions P, which may then be converted into normal deviates Z in the same manner in which p was converted into z. For these values of x we now put

$$f(x) = 100 + \sigma_f Z , \qquad (3)$$

where σ_f^2 is as given by equation (2). The mean and standard deviation of $f(x)$ will then be as previously specified. We may regard $f(x)$ as the I.Q. corresponding to a score x and an age given by $a = 0$, since in dealing with the total distribution of score we have in effect averaged out age. For values of x other than the end points of the grouping intervals $f(x)$ must be obtained by interpolation or extrapolation. In theory it might be possible to make some assumption regarding the mathematical form of $f(x)$. This approach to the problem has, however, not been followed owing to the difficulties which would arise in estimating the unknown parameters thereby involved.

To find the I.Q. y corresponding to any pair of values of x and a we use both equations (1) and (3). Hence

$$\begin{aligned} y &= f(x) - \beta a \\ &= (100 - \beta a) + \sigma_f Z . \end{aligned} \qquad (4)$$

It will be found convenient to calculate the I.Q.s y_1 corresponding to the lowest value of a required, say, a_1. Thus

$$y_1 = (100 - \beta a_1) + \sigma_f Z . \qquad (5)$$

For other values of a the I.Q. y is then given by

$$y = y_1 - \beta(a - a_1) . \qquad (6)$$

Numerical Example. To illustrate the method of standardization a numerical example is given below. The data represent the score distributions of a complete year-group for a Moray House group-test. The ages are given in years and completed months at date of test, and range from 10 : 8 (10 years 8 months) to 11 : 7.

Two values of x_0 were chosen, namely, 49·5 and 79·5 ; these correspond to the sets of values p, z and p', z' respectively. The slopes of the two lines found by the method of least squares differ remarkably little ; they are $-$·03524 and $-$·03514 respectively. We therefore estimate β as $15 \times$ ·0352 = 0·528. The value of σ_f is then given by $\sigma_f = \sqrt{225 + 12(0{\cdot}528)^2} = 15{\cdot}11$.

In constructing the conversion table it is inconvenient to use the Z values corresponding to the end points of the score intervals, namely, 9·5, 19·5, 29·5, etc. We therefore interpolate or extrapolate to obtain the values for $x = 10, 20, \ldots 90$. Using equation (5) we then find the I.Q.s y_1 for age 10 : 8. The difference between the lowest age 10 : 8 and the median age 11 : 1½ is 5·5 months.

TABLE I. DISTRIBUTION OF SCORE AND AGE

Score interval	Age in years and completed months at date of test												Totals	Cumul. Totals	*P*	*Z*
	10 : 8	10 : 9	10 : 10	10 : 11	11 : 0	11 : 1	11 : 2	11 : 3	11 : 4	11 : 5	11 : 6	11 : 7				
90–100	19	15	22	26	21	25	25	25	21	28	31	42	300	5368		
80– 89	31	40	36	30	45	37	47	34	39	51	46	60	496	5068	·944	1·589
70– 79	38 424	45 366	48 419	44 421	40 390	47 371	41 411	50 392	40 348	43 331	59 362	53 337	548	4572	·852	1·045
60– 69	38	36	55	52	48	67	53	52	41	49	54	48	593	4024	·750	·674
50– 59	46	37	46	57	51	53	58	47	45	44	50	39	573	3431	·639	·356
40– 49	57 302	39 248	41 270	60 268	41 251	44 204	59 259	48 243	47 222	47 195	38 199	34 197	555	2858	·532	·080
30– 39	53	41	63	45	65	29	47	59	45	39	40	44	570	2303	·429	−·179
20– 29	54	47	48	49	47	42	55	46	48	34	40	44	554	1733	·323	−·459
10– 19	64	60	56	57	54	45	49	46	46	40	49	40	606	1179	·220	−·772
0– 9	74	61	62	57	44	44	49	44	36	35	32	35	573	573	·107	−1·243
Totals ..	474	421	477	477	456	433	483	451	408	410	439	439	5368			
p	·637	·589	·566	·562	·550	·471	·536	·539	·544	·476	·453	·449				
z	·350	·225	·166	·156	·126	−·073	·090	·098	·110	−·060	−·118	−·128				
p′	·895	·869	·878	·883	·855	·857	·851	·869	·853	·807	·825	·768				
z′	1·254	1·122	1·165	1·190	1·058	1·067	1·041	1·122	1·049	·867	·935	·732				

TABLE II. CONVERSION TABLE

Age at Test Date	RAW SCORES																		
	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
10 : 8	80	85	88	92	94	96	98	100	102	104	106	109	111	113	116	119	123	127	133
10 : 9	79	84	88	91	94	96	98	100	102	104	106	108	110	113	115	119	122	127	132
10 : 10	79	83	87	90	93	95	97	99	101	103	105	107	110	112	115	118	122	126	132
10 : 11	78	83	87	90	92	95	97	99	101	103	105	107	109	112	114	117	121	126	131
11 : 0	78	82	86	89	92	94	96	98	100	102	104	106	109	111	114	117	121	125	131
11 : 1	77	82	86	89	91	94	96	98	100	102	104	106	108	111	113	116	120	125	130
11 : 2	77	81	85	88	91	93	95	97	99	101	103	105	108	110	113	116	120	124	130
11 : 3	76	81	85	88	90	92	95	97	99	101	103	105	107	110	112	115	119	124	129
11 : 4	76	80	84	87	90	92	94	96	98	100	102	104	107	109	112	115	119	123	129
11 : 5	75	80	84	87	89	91	94	96	98	100	102	104	106	109	111	114	118	123	128
11 : 6	75	79	83	86	89	91	93	95	97	99	101	103	106	108	111	114	118	122	128
11 : 7	74	79	83	86	88	90	93	95	97	99	101	103	105	108	110	113	117	122	127

Hence, putting $a_1 = -5{\cdot}5$, we find that

$$y_1 = 100 - 0{\cdot}528 \times (-5{\cdot}5) + 15{\cdot}11Z$$
$$= 102{\cdot}90 + 15{\cdot}11Z .$$

For other ages the I.Q. y is obtained by subtracting multiples of β, according to equation (6).

Fitting of straight line for $x_0 = 49{\cdot}5$

·350	−·128	·478	11	5·258
·225	−·118	·343	9	3·087
·166	−·060	·226	7	1·582
·156	·110	·046	5	·230
·126	·098	·028	3	·084
−·073	·090	−·163	1	−·163
·950	−·008	·958		10·078

Slope $= -10{\cdot}078/286 = -{\cdot}03524$.

Values of Z and I.Q.s for age 10 : 8

x	Z	x	Z	I.Q.* for 10 : 8 (y_1)
		95		(133·00)
89·5	1·589	90	1·621	127·39
		85		(122·82)
79·5	1·045	80	1·068	119·04
		75		(115·97)
69·5	·674	70	·691	113·34
		65		(110·84)
59·5	·356	60	·371	108·51
		55		(106·35)
49·5	·080	50	·093	104·31
		45		(102·35)
39·5	−·179	40	−·166	100·39
		35		(98·34)
29·5	−·459	30	−·444	96·19
		25		(94·03)
19·5	−·772	20	−·752	91·54
		15		(88·37)
9·5	−1·243	10	−1·215	84·54
		5		(80·00)

* Figures in brackets are interpolated or extrapolated.

In the conversion table I.Q.s have been given correct to the nearest whole number. Various checks may be made on the accuracy of the table : we may note, for example, that the median score of 46·4 and median age 11 : 1½ correspond to an I.Q. of 100, and that in each month-group the percentages of children with I.Q.s of over 100 and 115 (say) are near the expected values of 50 per cent. and 16 per cent. respectively.

REFERENCES

1. Thomson, G. H. (1932). 'The standardization of group tests and the scatter of intelligence quotients.' *Brit. Journ. Educ. Psychol.*, II, 92 and 125.
2. Pearson, K. (1930). *Tables for Statisticians and Biometricians*, Part I, 3rd ed., 1, Table I.

LINEAR AND NON-LINEAR DISCRIMINATING FUNCTIONS

By A. LUBIN

University College and Maudsley Hospital, London

I. *Problem.* II. *Mathematical Derivations.* III. *The Prediction of Qualitative Attributes from Qualitative Categories.* IV. *A Non-parametric Test of Significance.* V. *Summary.*

I. PROBLEM

The main purpose of this paper [1] is to present a summary of recent work in discriminant functions. Discriminant functions represent a fairly new addition to the statistical techniques that can be used by psychologists. Primarily they are solutions to problems in prediction. Given certain measures of physical or mental traits, can we predict that occupation in which an individual will succeed ? Can we state that therapy is most likely to benefit patients who have a certain pattern of scores on a personality test ? In these cases we wish to predict a qualitative attribute of an individual from known quantitative variables. The discriminant function is used whenever there are a set of mutually exclusive classifications and an individual must be assigned to one of these classes on the basis of a standard set of quantitative scores.

Historically speaking, the discriminant function was used long before it was given a name. In 1936, R. A. Fisher described a discriminant function as a linear combination of a set of quantitative variables, which is used to assign an individual to one of two mutually exclusive groups (11). But psychologists had used what Burt has called a ' multiple point biserial ' correlation for some time (3, 4, and refs.). The point biserial procedure was derived from the assumption that all individuals within one group should have the same score on the criterion. There was a widespread belief that such a procedure was illegitimate. If psychological measurements are sampled from the normal distribution, the assumption of a point distribution becomes untenable. But definitive justification of the ' multiple point biserial ' correlation appeared in the literature when it was shown that the regression weights given by the multiple biserial (which assumed a normal distribution of the criterion), the multiple point biserial (which assumed a point distribution), and the discriminant function of R. A. Fisher were all mutually proportional (4, 5, 30).

Properly to appreciate the importance of the discriminant function we must place it in its setting among other prediction procedures. Table I illustrates four possible types of simple prediction problems.[2] In the last line the problem is raised of how to predict from one set of qualitative variates to another. Although there are special solutions for this problem in terms of quantifying the qualities, the general solution has not yet been stated in purely qualitative terms. One solution is considered in Section III, pp. 99–100.

[1] As will be seen, the writer is indebted to Professor Sir Cyril Burt for many of the ideas in the article. Acknowledgement must also be made to Dr. M. Hamilton, Mr. A. Jonckheere, Mr. A. Summerfield, Miss D. Webb, and my colleagues at University College and Maudsley Hospital who have been invaluable sources of stimulation and instruction.

[2] Somewhat similar classifications will be found in 4 and 16 : cf. also this *Journal*, I, pp. 139–40.

TABLE I

Dependent Variate	Independent Variate(s)	Method of Solution
1. Quantitative	Quantitative	Multiple Regression
2. Quantitative	Qualitative	Analysis of Variance
3. Qualitative	Quantitative	Discriminant Function
4. Qualitative	Qualitative	?

Probably most users of analysis of variance have never thought of it as a method of predicting a quantitative variable from one or more qualitative variables. However, the simple probability models used in setting up the equations of analysis of variance postulate a quantitative effect for each category of classification. They can, therefore, be regarded as prediction equations.

Placing discriminant functions in the same table as analysis of variance and multiple regression implies that the discriminant technique is of the same order of importance. In the writer's opinion this is true. The field of vocational and educational guidance alone should find the discriminant function an indispensable tool. In all fields of empirical enquiry, there will be a shift of emphasis. We will no longer ask only the question, "Are the differences between the groups statistically significant?" The new questions will be, "Are these differences of any practical use? Can they be used to allocate individuals to their proper classification?"

II. MATHEMATICAL DERIVATIONS

The Prediction of Quantitative Variates from Qualitative Attributes. The simplest situation in analysis of variance is the one-way classification. As a prediction problem, it is a question whether a single quantitative attribute, such as intelligence, can be predicted from a single qualitative attribute, such as membership in a political party. The basic equation is:

$$x_{ij} = A_j + z_{ij}, \qquad (1)$$

i.e., the observed score, x, of the ith individual in the jth class is equal to the overall level of the jth class (A_j) plus some positive or negative error effect. Thus A_j is the best estimate of the score that can be made from knowing that the individual comes from the jth class. The error of prediction is z_{ij}. The least square estimate of A_j will obviously be the mean of the n_j individuals in the jth class (which will be denoted by $\bar{x}_{.j}$). We will therefore assign to each member of the class j the score $\bar{x}_{.j}$.

The attribute of class membership has in effect been quantified by equation (1). What was a series of classes can now be represented by a series of points on a continuum represented by the x variable. It is this metricization of the qualitative attribute of class membership that makes the linear discriminant function possible. We shall later derive the linear discriminant function from this type of reasoning. It will be important to remember that equation (1) is essentially an estimation function. It gives a predicted score ($x_{.j}$) for each individual which minimizes the sum of squared discrepancies.

The Pearson product-moment correlation between the predicted score $\bar{x}_{.j}$ and the observed score x_{ij} will equal the ratio of the standard deviation of the predicted score to the standard deviation of the observed score. This correlation is the correlation ratio, *eta*. And we may recall that giving the mean of each class as the predicted score for that class maximizes the correlation ratio. (See 16, pp. 264–8.)

We shall now give the usual analysis of variance equations leading up to the F ratio. These equations will be used later in the derivation of the linear discriminant function.

Substituting the least-square estimate for A_j we can write from (1) that $z_{ij} = x_{ij} - \bar{x}_{.j}$. Subtracting the general mean $\bar{x}_{..}$ from both sides of equation (1) and substituting for z_{ij}, we get the standard form

$$(x_{ij} - \bar{x}_{..}) = (\bar{x}_{.j} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{.j}) \qquad (2)$$

The deviation of the observed score from the general mean equals the deviation of the group mean from the general mean plus the deviation of the observed score from the group mean. If we square each term for each individual and sum over the n_j individuals in each group and then over the c groups, we find the well-known important result that

$$\sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{..})^2 = \sum_{j=1}^{c} n_j(\bar{x}_{.j} - \bar{x}_{..})^2 + \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{.j})^2 \qquad (3)$$

The deviance about the general mean equals the deviance between group means plus the deviance within groups. The equation for the square of the correlation ratio can be written as

$$R^2 = \frac{\text{deviance between groups}}{\text{total deviance}} . \qquad (4)$$

The equation for the F ratio can be written as a function of R^2

$$F = \frac{N - c}{c - 1} \cdot \frac{R^2}{1 - R^2} , \qquad (5)$$

which shows the basic relation between the two statistics.

Derivation of the Linear Discriminant Functions as Canonical Variates. We can now use the above five equations to metricize qualities and derive a linear function which will best predict the quantized attributes. The derivation has been given by R. A. Fisher for the discrimination of two groups (11). An extension to more than two groups was suggested by Burt for psychological and educational purposes (4); and (with a different line of approach) has also been described in an article by C. R. Rao (23). Rao's somewhat brief derivation was given in terms of maximizing Mahalanobis's D^2. Since the D^2 statistic is unfamiliar to most psychologists, the derivation here (which is similar to that of Burt) will be in terms of maximizing R^2, which will also maximize the F ratio.

The F ratio (or its equivalent, the correlation ratio) tells us if class membership significantly differentiates the individual with respect to the quantitative variable. We will now invert the problem of predicting quality from quantity, and assign numbers to each category from the quantitative variables in such a way as to maximize the F ratio.

The question is put in this manner: Given q test scores for the ith person in the jth class, how can we form the linear combination of the q scores that makes R^2 a maximum?

Let y be the best linear function and let $u_1, u_2, \ldots u_q$ be the best q weights. The notation will be: t represents the tests and runs from 1 to q; j represents the groups and runs from 1 to c; i represents the individuals and runs from 1_j to n_j within each group. $N = \sum_{j=1}^{c} n_j$. Then the y score for each individual will be the sum of all his x scores with each x weighted by the appropriate u.

$$y_{ij} = u_1x_{1ij} + u_2x_{2ij} + \ldots u_tx_{tij} + \ldots u_qx_{qij} = \sum_{t=1}^{q} u_tx_{tij} . \qquad (6)$$

In matrix notation:

$$y_N = X_{ij}u ,$$

where y_N is a column vector of N elements, one for each person; X_{ij} is the raw score matrix of N individuals for q tests, whose general element is x_{tij}; and u is the column vector of weights, one for each of the q tests, which will specify the best linear discriminant function. N is assumed to be greater than q.

Let us substitute for y_N in the analysis of variance equations and state the formula for the squared correlation ratio, R^2, in terms of $X_{ij}u$, and matrices derived from X_{ij}. Thus $(y_{ij} - \bar{y}..) = (y_{ij} - \bar{y}_{\cdot j}) + (\bar{y}_{\cdot j} - \bar{y}..)$ in matrix notation becomes

$$Xu = X_i u + X_j u , \tag{7}$$

where X is an N by q matrix whose general element is $(x_{tij} - \bar{x}_{t\cdot\cdot})$, the deviation of the observed score on test t from the general mean.

X_i is an N by q matrix whose general element is $(x_{tij} - \bar{x}_{t\cdot j})$, the deviation of the observed score on test t from the mean of the jth group.

X_j is an N by q matrix whose general element is $(\bar{x}_{t\cdot j} - \bar{x}_{t\cdot\cdot})$, the deviation of the group mean from the general mean.

u is the set of weights defining the linear discriminating function. In equation (3) the sum of squares or deviance about the general mean was equal to the deviance between group means plus the deviance within groups. Writing the same equation for y in matrix notation :

$$u'X'Xu = u'X_i'X_i u + u'X_j'X_j u . \tag{8}$$

Our notation will be simplified if we use the following definitions.

1. $X'X \equiv G$, the " general " dispersion or codeviance matrix. The general element in the sth row and the tth column is

$g_{st} \equiv \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{tij} - \bar{x}_{t\cdot\cdot})(x_{sij} - \bar{x}_{s\cdot\cdot})$. It is a square symmetric matrix of order q, as are the two matrices defined below, W and B. The element $g_{st} = N r_{st}\, \sigma_s\, \sigma_t$.

2. $X_i'X_i \equiv W$, the " within groups " codeviance matrix.

$$w_{st} = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (x_{sij} - \bar{x}_{s\cdot j})(x_{tij} - \bar{x}_{t\cdot j}).$$

3. $X_j'X_j \equiv B$, the " between groups " codeviance matrix.

$$b_{st} = \sum_{j=1}^{c} n_j (\bar{x}_{s\cdot j} - \bar{x}_{s\cdot\cdot})(\bar{x}_{t\cdot j} - \bar{x}_{t\cdot\cdot}).$$

These three q by q codeviance matrices are analogous to the single variable deviances in equation (3). For example $G = W + B$.

The formula for the square of the correlation ratio of the discriminating function, y, is similar to equation (4)

$$R^2 = u'Bu/u'Gu . \tag{9}$$

The formula for R^2 will now be used to derive the linear discriminating function. We shall seek the vector u which makes R^2 a maximum. The procedure will be to differentiate R^2 with respect to u, set the derivative equal to zero and solve the resulting equation for u. We obtain

$$\frac{dR^2}{du} = \frac{(u'Gu)\, 2u'B - (u'Bu)\, 2u'G}{(u'Gu)^2} = 0. \tag{10}$$

Multiply equation (10) by the scalar $(u'Gu)$, substitute $R^2 = u'Bu/u'Gu$ from equation (9), simplify, and the result is

$$u'B - R^2\, u'G = u'\,(B - R^2 G) = 0. \tag{11}$$

Post-multiplying by G^{-1}

$$u'\,(BG^{-1} - R^2 I) = 0. \tag{12}$$

This, in slightly different notation, is the equation reached by Burt (4). As he has suggested, it can also be derived direct from the ordinary expression for a canonical correlation, and can readily be extended to the multidimensional case by means of a preliminary factor analysis (4 ; 6, p. 105).

Equation (12), and the conditions of the problem, tell us that the u we seek is the latent vector corresponding to the largest latent root of BG^{-1}. This can be found by the usual iterative methods (4). BG^{-1} is non-symmetric, and one must be careful to *pre-multiply* by the trial row vector. The transposed matrix $G^{-1}B$ must be used if *post-multiplication* by a trial column vector is desired. The largest latent root is the largest R^2 obtainable.

Equation (12) also implies that there are as many latent roots which are solutions as there are linearly independent rows or columns of $B\,G^{-1}$. The number of independent linear discriminating functions is therefore equal to the rank of $B\,G^{-1}$. There are definite limits to the rank of $B\,G^{-1}$. G must be a non-singular matrix, i.e., must be of rank q, or else the inverse G^{-1} will not exist. On the other hand, B has a maximum rank of $c - 1$. This is a result of the fact that the sum of any column of X_j equals a constant, zero. This means that the maximum number of independent rows in X_j is $c - 1$. So if q is greater than $c - 1$, the rank of $B\,G^{-1}$ is $c - 1$. If $c - 1$ is greater than q, the rank of $B\,G^{-1}$ is q.

While obviously the best single linear discriminating function is that one which is associated with the largest latent root, the set of all linear discriminating functions defines the number of dimensions which the groups occupy in q dimensional space.

Multivariate linear discriminating analysis can be likened to a form of factor analysis which extracts only those factors which help to distinguish groups from one another. I therefore suggest using the phrase "canonical variates" to denote linear discriminating functions extracted in this manner. Any linear combination of scores on which the group means tend to be equal will be discarded automatically. This raises the possibility of experimental verification of postulated factors. If a factor can be defined in terms of groups of homogeneous individuals, then the discriminating function which maximizes the discrimination between these groups satisfies the definition of the postulated factor.

The set of latent vectors associated with $B\,G^{-1}$ will form a q by r matrix which we can call U. U cannot be compared directly with the usual factor analysis matrix F. F is usually used to denote the set of correlations of q tests with r uncorrelated factors. It is therefore the set of regressions of the tests on the factors. U is the set of regressions of the r discriminating functions on the q tests when the standard deviations of the tests have not been made equal to unity. F can be turned into a matrix of regressions of factors on tests by one of several formulæ. $R^{-1}\,F$ is one such formula (where R is the q by q correlation matrix). When multiplied by the appropriate diagonal matrix of standard deviations, these factor regression matrices will be directly comparable to U. Tests of significance for hypothetical discriminating functions versus observed discriminants have been discussed by Rao (24) and Fisher (14). But special procedures will have to be worked out for verifying factors found in previous factor analysis studies by comparing U with F.

It is only permissible to define a factor operationally in terms of a linear discriminating function, when that factor has itself been defined in terms of a discrimination between groups. H. J. Eysenck, however, in his studies of the ' dimensions of personality ' speaks of factors which are defined in terms of their power to differentiate between various diagnostic types of mental disorder. In a later article I hope to explain my view more fully by means of illustrative examples, using (amongst other material) data drawn from Dr. Eysenck's own work.

Multivariate discriminating analysis can also be used to answer the question of whether a set of categories really form a single linear continuum. If the rank of U is greater than one, then the groups cannot be represented as points or intervals along a single scale. The best linear discriminating function would still not suffice to represent the differences between the groups. For example: discriminating function analysis could be used to test if political groups such as Liberal, Conservative, and Socialist actually fall on a single dimension of Radicalism–Conservatism.

When the covariance matrices of all groups are assumed equal, the best discriminating functions are all linear in form. A full discussion of this case with a worked example is given by Rao and Slater (25). A further discussion of this point will be found in Section II (b) of this paper.

The Prediction of Qualitative Attributes from Quantitative Variables by the Maximum Likelihood Discriminating Function. The underlying weakness in the usual approach to discriminating functions is that unless the categories lie on a single dimension they cannot be adequately discriminated from one another. Quantifying the categories may force an actual multidimensional set of qualities on to a single straight line. The procedure may really be distorting the data and obscuring real differences between the categories.

The maximum likelihood discriminating function developed by Rao avoids this difficulty (23). In this method, the pattern of scores is used to discriminate between groups of individuals without the step of condensing each individual's set of scores into a point on a linear continuum. This basically different approach releases the psychologist from the necessity of forcing quantification on essentially qualitative data.

A full discussion of the mathematical basis for the method will be found in the paper by Rao. More elementary discussions will be found in the papers by Welch (29) and C. A. B. Smith (27).

Smith applied the maximum likelihood method to the problem of discriminating two groups. His solution was a likelihood ratio which, interestingly enough, simplifies to Fisher's linear discriminating function when the variance–covariance matrices of the two groups are equal.

Briefly stated, the method consists of calculating the likelihood for each individual with respect to each group. The individual is then assigned to the group for which he has the maximum likelihood. Rao's full method would lead to as many likelihood scores for each individual as there were groups. Smith's solution was to calculate the likelihood ratio, the ratio of the likelihood of belonging to one group to the likelihood of belonging to the other group. When there are only two groups, this is a single score. It is possible to examine the logical basis of the maximum likelihood solution without going deeply into the mathematics of the problem. Smith gives a simple explanation of the two-group case. In the following we shall consider the simplest case that will illustrate the full non-linear discriminating method, namely, the case of three groups. Let us, however, begin by glancing at the *univariate* or linear case.

Given three groups of people, Group A, Group B, and Group C, who have taken a single test X, how can we allocate each individual to a group, merely from a knowledge of his score on X, so as to minimize the percentage of misclassifications ? The logical structure of this question is identical with the problem of quantitative prediction. A set of mutually exclusive groups or categories is a qualitative variable just as a set of scores is a quantitative variable. Allocation to a group is equivalent to prediction of a quantitative score. And the concept of minimizing the percentage of misclassified individuals is of the same importance and plays the same role as the least squares concept of minimizing the squared errors of prediction. (It will be recalled that the canonical linear discriminating function described in Section I was derived from least square considerations.)

A graph of the three groups is given in Fig. 1. x_1, x_2, and x_3 are individuals possessing scores on X at the point indicated. The ordinate L_{c1} represents the number of Group C individuals that may be expected to have the score x_1. Similarly, the ordinates L_{a1} and L_{b1} represent the number in Groups A and B that will have the score x_1.

We cannot take any specific individual who has the score x_1 and say that these frequency elements correspond to the probabilities that he has come from these various groups. Since we have specified the membership of each individual at the beginning of our problem, the probability that any individual belongs to a specific group is either zero or unity.

However, the concept of ' likelihood ' as used by R. A. Fisher may be used here. The three ordinates represent the three likelihoods that a person with score x_1 will belong to each of the three groups. Rao has demonstrated the important theorem that if each person is assigned to the group for which his likelihood is greatest, the number of misclassifications is a minimum. In this case all persons having the score x_1 would be assigned to Group A. Persons with scores x_2 and x_3 would be assigned to Group B and Group C respectively. If there were equal numbers in each group it is obvious that this would lead to a minimum number of misclassifications. If the number of people in the universe differs from the sampled number in the group, then each frequency element must be weighted by an amount proportional to the number in the group for this universe. In what follows it will be assumed that the proportions of the three groups in the universe are equal.

It will be noted that there is an interval, a, along the X axis where La is always larger than Lb or Lc. Similarly there is an interval b where the likelihood of belonging to Group B is greater than the other likelihoods. And for all scores in the interval c, the likelihood is always greatest for Group C. Rao's theorems state that such regions exist and can always be found if the ordinate for each score in each group is known.

Now the point which separates the regions is the point where the likelihood functions are equal, i.e., the point where the curves intersect. If we knew the distribution function for Group A, then the number of individuals from Group A who would be misclassified could be calculated. The number misclassified as Group B is equal to the integral of La over the interval b. Similarly the number of Group A individuals misassigned to Group C is the area under the A frequency curve for the interval c. It will be noted that as yet we have not assumed any specific type of distribution curve ; and the above statements are true no matter what form the distribution function takes.

As Burt has shown (see Appendix, p. 104), the point X where two normal curves intersect is given by the formula

$$X = \frac{(\sigma_1^2\bar{x}_2 - \sigma_2^2\bar{x}_1) \pm \sigma_1\sigma_2\sqrt{\left(\bar{x}_1 - \bar{x}_2\right)^2 + 2\left(\sigma_1^2 - \sigma_2^2\right)\left[\ln \frac{N_1\sigma_2}{N_2\sigma_1}\right]}}{\sigma_1^2 - \sigma_2^2}$$

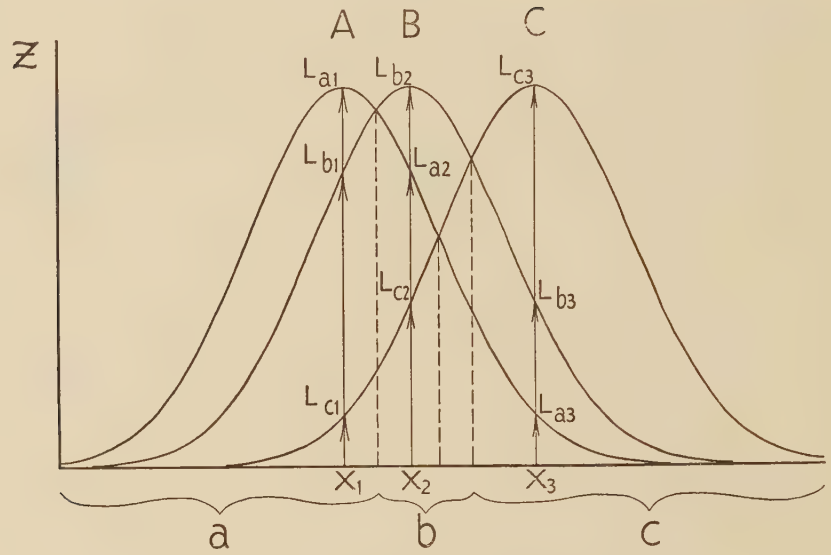

FIG. 1. Discrimination : the one-dimensional problem.

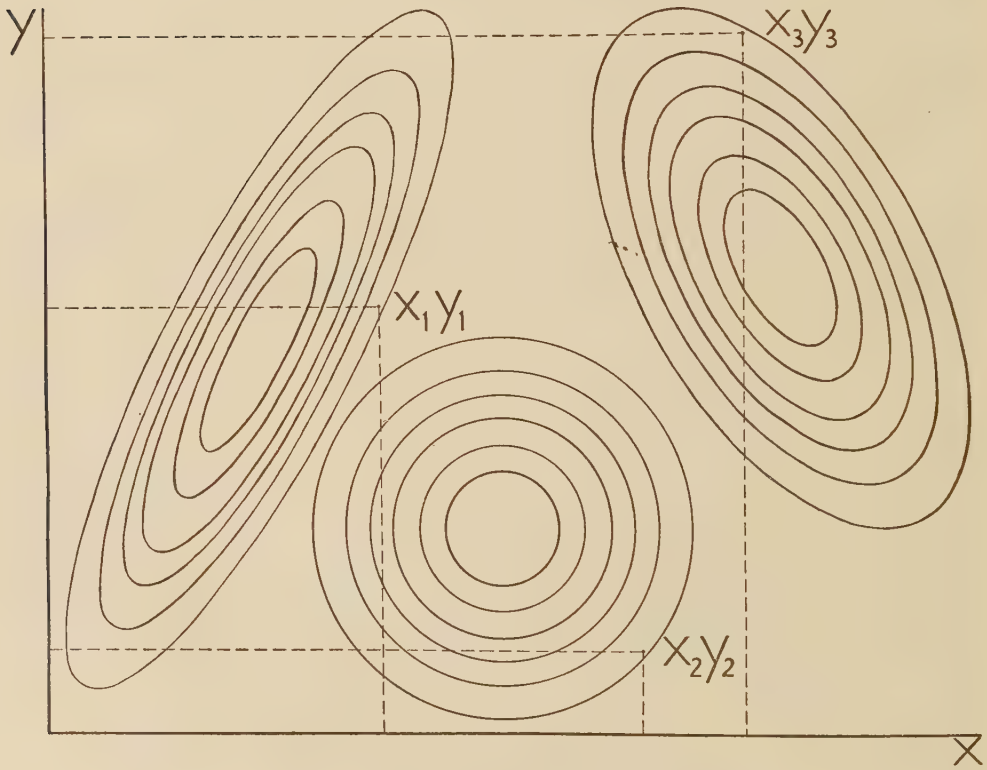

FIG. 2. Discrimination : the two-dimensional problem.

where $X \equiv$ the raw score at the point of intersection, $\bar{x}_1 \equiv$ the mean of the first group, $\bar{x}_2 \equiv$ the mean of the second group, $\sigma_1 \equiv$ the standard deviation of the first group, $\sigma_2 \equiv$ the standard deviation of the second group, $N_1 \equiv$ the number in the first group, $N_2 \equiv$ the number in the second group. When the N's and σ's are equal the intersection point is midway between the two means. Burt, who first obtained the above equation (2), seems to have been the first to propose the point of intersection as the cut-off point for allocation of individuals into groups.[1]

Let us now consider the bivariate case. We can plot each person on a two-dimensional graph with x and y as the axes. If we think of frequency as a third dimension orthogonal to the plane of the graph, then each group resembles a hill. The height of the hill represents the relative frequency with which a particular pair of scores x_1y_1, x_2y_2, etc., is found in the group. Fig. 2 is a rough schematic representation of the situation.

In Fig. 2 contour lines have been used to indicate height, i.e., the frequency element. It is obvious that there are regions where all individuals have a greater likelihood of belonging to one group than the other two. Rao has outlined methods of obtaining these regions (23). For our purposes, the fact that they exist is sufficient.

In this bivariate case it would be necessary to compute the three likelihoods for each pair of scores and assign each person to the group where his likelihood is greatest. The number of misclassifications resulting from this procedure is always equal to or smaller than the misclassifications that would arise from using only one variable.

The same procedure can easily be extended to the case of more than two variables. For each person, the three likelihoods (one for each group) would be calculated for each set of scores. Allocation would be to the group with the highest likelihood. It is clear that increasing the number of groups merely increases the number of likelihoods that must be calculated. The maximum likelihood discriminating technique is, then, applicable to any number of groups. Increasing the number of quantitative variables increases the complexity of the likelihood functions but does not increase the number of likelihoods that must be computed.

A real difficulty arises in the actual calculation of the likelihoods. The number of individuals in any group is, of course, finite. Empirical frequencies could be determined for grouped values of one variable. But the task would soon become impossible if we attempted to determine the relative frequencies of each pair or triplet of scores. To actually determine these proportions empirically would demand an impossibly large number of people in the sample. This same difficulty arises in the prediction of qualitative attributes from other qualitative variables. It is discussed in Section III.

The only practical procedure appears to be an assumption about the theoretical nature of the distribution within each group. Special considerations may lead to the selection of a binomial, a rectangular, a Poisson, or other frequency distribution as best representing the shape of the distribution curve within each group. In many, if not most, cases the normal multivariate function will prove most convenient.

To use a theoretical frequency distribution it will be necessary to calculate, for each group, the constants which define the function. In the case of a normal univariate distribution, the mean and standard deviations are sufficient. The score of any individual may then be substituted into the equation for each group. Therefore each individual will have as many likelihoods as there are groups.

The equation for the multivariate normal distribution will involve the calculation of all means, standard deviations, and inter-correlations. The inverse of each matrix of inter-correlations must be computed. The likelihood function for each person will involve a generalized quadratic function whose coefficients are the elements of the inverted covariance matrix. This process has been described in detail by C. A. B. Smith (27). Since Smith is concerned only with the two-group case, he uses as his discriminant function the ratio of the two likelihoods, or, rather, twice the natural logarithm of the likelihood ratio. It will be recalled that for the multiple-group case it is necessary only to select the largest likelihood. Even so the likelihood function for the multivariate normal distribution is likely to be simplified if logarithms instead of the actual frequencies are used. Another aid to simplification is to reduce the quantitative variables which are being used as predictors to as few orthogonal variables as possible. The canonical reduction connected with the best linear discriminating function (see equation (12)) seems likely to furnish the optimum set of predictors. But this elaborate procedure is unworkable for a large number of variables. Some simplified form of *a priori* factor analysis such as that suggested by L. S. Penrose (20) might prove convenient and relatively efficient. The condensed set of variables should be so selected as to be linearly independent

[1] *Mental and Scholastic Tests* (1921), p. 165. The conditions of the problem assume that we choose the point of intersection *between* the two means, i.e., take the sign before the radical to be positive. In certain exceptional cases the relative size of σ_1 and σ_2 or of N_1 and N_2 may make this impossible. Note that, here and below, I use "ln" as an abbreviation for "natural logarithm of," i.e., for "$\log_e$."

since the multivariate normal distribution requires the inverse of the correlation matrix. This inverse will not exist if the variables are not linearly independent of one another.

(a) *The Non-Linear Discriminating Function for the Case of a Multivariate Normal Distribution.* The multivariate normal distribution is of special interest because so many psychological variables are distributed normally. When the variables used are themselves the sum of a large number of variables, it is very likely that these sums are distributed approximately as the normal curve. This follows from the central limit theorem (9, p. 232). It is possible in the psychological field to construct variables that closely approximate a normal distribution of quantitative scores by obtaining the sum of a large number of repeated measures of the same trait. The discriminating equation will accordingly be derived assuming that the distribution for each category or group is multivariate normal.

A matrix notation will conveniently be used. To distinguish vectors from scalars, the letters denoting the variable elements will be underlined when they represent vectors. Let $C_j \equiv$ the $q \times q$ covariance matrix of the jth group, $\underline{x}_i \equiv$ the 1 by q row vector of observed raw scores for the ith individual, $\underline{d}_{ij} \equiv \underline{x}_i - \underline{m}_j \equiv$ the 1 by q row vector of the q scores expressed as deviations from the q means of the jth group; $\underline{d}_{ik} = \underline{x}_i - \underline{m}_k$, the same raw score vector expressed as a deviation from the vector of means for the kth group. And let $n_j \equiv$ the number of individuals in the jth group; $N \equiv \sum_{j=1}^{c} n_j$; $p_j \equiv n_j/N$; and $f_j(\underline{x}_i) \equiv$ the frequency of the set of observed raw scores $\underline{x}_i$ in group j.

There will be c discriminating functions, one for each group. The standard operating procedure would be to calculate $f_j(\underline{x}_i)$ for all the c groups and then assign all individuals having the pattern of scores, $\underline{x}_i$, to the group where the frequency of $\underline{x}_i$ is highest. This will minimize the total number of misclassifications.

The equation for the frequency of $\underline{x}_i$, assuming a multivariate normal distribution, is:

$$f_j(\underline{x}_i) = \frac{n_j}{(2\pi)^{q/2} \mid C_j \mid^{\frac{1}{2}}} \exp\left[-\tfrac{1}{2}\, \underline{d}_{ij}\, C_j^{-1}\, \underline{d}_{ij}'\right] . \tag{13}$$

Similarly the frequency for $\underline{x}_i$ in the kth group would be:

$$f_k(\underline{x}_i) = \frac{n_k}{(2\pi)^{q/2} \mid C_k \mid^{\frac{1}{2}}} \exp\left[-\tfrac{1}{2}\, \underline{d}_{ik}\, C_k^{-1} \underline{d}_{ik}'\right] . \tag{14}$$

Thus the frequency of any pattern of scores, $\underline{x}_i$, can be calculated for each of the c groups. We can simplify equation (13) somewhat. The term $(2\pi)^{q/2}$ is a constant for all calculated frequencies for all observed values of $\underline{x}_i$, and does not need to be considered.

Let $$L_{ij} \equiv +2 \ln [f_j(\underline{x}_i)] + q \ln 2\pi. \tag{15}$$

Then $$L_{ij} = 2 \ln(n_j) - \ln \mid C_j \mid - \underline{d}_{ij}\, C_j^{-1} \underline{d}_{ij}' . \tag{16}$$

L_{ij} is a quadratic expression which can be used in place of the frequency. When the frequency of $\underline{x}_i$ is near zero, L_{ij} will tend to minus infinity. When the frequency of $\underline{x}_i$ is unity or greater than unity, L_{ij} will be positive.

There is still another form of the discriminating function that is useful when the absolute number of individuals in each group is not known, but the relative proportions are given.

Let $$\lambda_{ij} \equiv 2 \ln \left[\frac{f_j(\underline{x}_i)}{N}\right] + q \ln 2\pi . \tag{17}$$

Then $$\lambda_{ij} \equiv 2 \ln p_j - \ln \mid C_j \mid - \underline{d}_{ij}\, C_j^{-1} \underline{d}_{ij}'. \tag{18}$$

Where $$p_j \equiv n_j/N . \tag{19}$$

λ_{ij} will vary from minus infinity to the maximum positive quantity of $(q \ln 2\pi)$.

Thus a personnel officer may not know what total number of bombardiers, navigators, and pilots will be required. If, however, he knows what the relative proportions should be in the Air Force as a whole, he can use λ_{ij}. This will allocate all available individuals to air crew training schools with the smallest number of misclassifications.

We have thus shown that, when the distribution within each group is multivariate normal, the best discriminating formula for each group is a quadratic expression of the q scores for each individual.

(b) *The Linear Discriminating Function for the Case of a Multivariate Normal Distribution.*

Let us assume that $$C_j = C_k = \ldots = C . \tag{20}$$

Then $$\lambda_{ij} = 2 \ln p_j - \ln \mid C \mid - (\underline{x}_i - \underline{m}_j)\, C^{-1} (\underline{x}_i - \underline{m}_j)'. \tag{21}$$

The quantity $(-\ln \mid C \mid)$ will be constant for all groups.

$$\lambda_{ij} + \ln \mid C \mid = 2 \ln p_j - \underline{x}_i C^{-1} \underline{x}_i' + 2\underline{x}_i C^{-1} \underline{m}_j' - \underline{m}_j C^{-1} \underline{m}_j' . \quad (22)$$

In equation (22) it can be seen that the quantity $-(\underline{x}_i C^{-1} \underline{x}_i')$ is a constant and does not vary over the c groups. It is now possible to define a new term, l_{ij}, which will be directly related to the likelihood of $\underline{x}_i$ in group j; but will involve only linear functions of $\underline{x}_i$

$$l_{ij} = \tfrac{1}{2}(\lambda_{ij} + \ln \mid C \mid + \underline{x}_i C^{-1} \underline{x}_i') = \underline{x}_i C^{-1} \underline{m}_j' + \ln p_j + \tfrac{1}{2}\underline{m}_j C^{-1} \underline{m}_j' . \quad (23)$$

Equation (23) states that l_{ij} is calculated for any $\underline{x}_i$ by applying the q weights given by $C^{-1} \underline{m}_j'$ and adding the constant term, $\ln p_j - \frac{1}{2} \underline{m}_j C^{-1} \underline{m}_j'$. When there are only two groups, l_{ij} can be shown to be directly proportional to the point-biserial multiple regression function, or to Fisher's linear discriminating function. Rao and Slater give a worked example for three variables and six groups (25) in which they use a function called L_i. In this article $l_{ij} \equiv L_i$.

Summary. Section II (a) shows that for a multivariate normal distribution in each group, and unequal covariance matrices, a quadratic function of the observed scores is the best discriminating function. Two such quadratic functions have been derived, L_{ij} and λ_{ij}, where

$$L_{ij} = 2 \ln [f_j (\underline{x}_i)] + q \ln 2\pi = 2 \ln n_j - \ln \mid C_j \mid - \underline{d}_{ij} C_j^{-1} \underline{d}_{ij}' \quad (24)$$

and

$$\lambda_{ij} = 2 \ln \left[\frac{f_j (\underline{x}_i)}{N}\right] + q \ln 2\pi = 2 \ln p_j - \ln \mid C_j \mid - \underline{d}_{ij} C_j^{-1} \underline{d}_{ij}' . \quad (25)$$

L_{ij} is to be used where the number of subjects, n_j, to be allocated to each of the c groups is known. λ_{ij} is for use when only the percentage of subjects, p_j, to be allocated to each of the c groups is known.

Section II (b) shows that, for a multivariate normal distribution in each group and a constant covariance matrix, C, a linear function l_{ij}, of the observed scores, is the best discriminating function.

$$2\, l_{ij} = 2 \ln \left[\frac{f_j \underline{x}_i}{N}\right] + q \ln 2\pi + \ln \mid C \mid + \underline{x}_i C^{-1} \underline{x}_i'$$

$$\text{or } l_{ij} = \underline{x}_i C^{-1} \underline{m}_j' + \ln p_j - \tfrac{1}{2}\underline{m}_j C^{-1} \underline{m}_j' . \quad (26)$$

A consideration of the equations for L_{ij}, λ_{ij}, and l_{ij} will show at once that class membership, as defined in Section II, is not at all an invariant relationship. Whether a recruit is classified as a bombardier or a pilot, by a selection officer, is likely to vary as the quotas for these jobs vary. If an invariant criterion of class membership is wanted, it will be necessary to set a confidence level for rejection from each group. L_{ij} or the other two equations can then be solved for the critical value which represents this confidence level. All individuals falling below this critical value will be rejected from the group.

Recently, there has been a great deal of interest and research on the problem of using test scores in differential diagnosis of mental disorders. Rapaport (26), Jastak (18), Rabin (21), and many others have advocated the use of certain difference scores and/or ratios of one score to another to allocate patients to various psychiatric categories. In particular, Jastak has concluded that ratio scores are most appropriate.

Obviously, this is the very type of problem for which the discriminating function was invented. From our previous discussions it should be clear that whenever the distribution of scores is multivariate normal (or can be satisfactorily approximated by a multivariate normal distribution) no ratio score can be an optimal method of discrimination. A difference score can be effective only if the variances and covariances are constant over the groups. The optimal discrimination will be achieved by using a quadratic function of the scores.

III. THE PREDICTION OF QUALITATIVE ATTRIBUTES FROM QUALITATIVE CATEGORIES

In Table I above, where various types of prediction problems were tabulated, the question of predicting qualities from other qualities was stated to have no known

solution. Actually many statistical techniques have been devised which deal with contingency tables, the basic data for this type of prediction problem. But all of these techniques test the deviation from a chance distribution, or measure the amount of association in a manner comparable to the correlation coefficient. What is needed is a workable solution to the question of how to predict class membership from knowledge of membership in a set of other classes.

Suppose a population of N individuals is divided into categories $A_1, A_2, \ldots A_j, \ldots A_p$ which we wish to predict. These will be called the criterion categories. Suppose we are given the further information that each person has a secondary classification in categories $B_1, B_2, \ldots B_i, \ldots B_q$ which are to be used to predict the criterion categories. These will be called the predictor categories.

We can now make out a $q \times p$ matrix whose general element is n_{ij}, the number of people who are in both the row B_i and the column A_j. The principle of maximum likelihood gives a direct simple solution for the best prediction of A from B. For any specified row B_i, find the n which is the largest in the row of values, $n_{i1}, n_{i2}, \ldots n_{ij}, \ldots n_{ip}$. If n_{i1} is the largest frequency in the B_i row, then the best prediction from B_i is to A_1. Thus a particular A category will be predicted from each B category. More than one B category will sometimes predict the same A category. There is no necessity for each B category to be linked with a different A category. It is easy to see that this method of taking the A category which has the highest frequency in a particular row and making it the prediction for that B row results in a set of predictions that minimizes the total number of misclassifications. This is the simple situation that prevails in the case of the single predictor variable.

In the case of two predictor variables, we are given the additional information that the individual has a tertiary classification on categories $C_1, C_2, \ldots C_h, \ldots C_r$. This leads to a three-dimensional table of $q \times p \times r$, whose general element is n_{hij}, the number of people who have a joint membership in C_h, B_i, and the criterion category A_j. The maximum likelihood method for the case of two predictors is formally the same as the case of prediction from B alone. For each combination C_h, B_i, find the prediction category A_j which has the largest frequency. When this category A_j is taken to be the prediction for an observed joint membership in C_h and B_i, again a minimum number of misclassifications will result.

But the number of possible combinations of B and C is qr. As the number of discriminators is increased arithmetically, the number of combinations increases geometrically. Since any sample is finite, as the number of predictors is increased, a point will soon be reached where the predicted A category from a particular subclass of the discriminators will be indeterminate. There will be no individuals who fall into this particular subclass.

Another difficulty is that the predictions must be established by a detailed investigation of empirical frequencies in each case. There seems to be no method whereby a set of parameters can be calculated which, when applied to the observed set of predictors, will give the best prediction of A.

Each contingency table could be first tested by the usual chi-square technique to ensure that the observed frequencies are unlikely to have arisen by chance. Coefficients of association can be computed for contingency tables. But the best measure, in view of the method we are using, is the percentage of incorrectly classified individuals.

R. A. Fisher has presented an interesting technique in his *Statistical Methods for Research Workers* (15) for quantizing the qualitative variates in a two-way $q \times p$ contingency table. Scores are assigned to categories A and B in such a way that the Pearson product-moment correlation between the quantized variates is a maximum. The extension to the multivariate case does not seem obvious. The basic difficulty with this approach is that it assumes exactly what we wish to avoid—that each category is a point on a linear continuum.

On the whole, the solutions presented here seem inadequate to the writer. Perhaps the presentation of the problem in this *Journal* will focus attention on the question and reveal that some existing technique can be used.[1] Most responses to situations are easier to classify in a qualitative rather than a quantitative manner; and the psychologist would find a great deal of use for a technique that required no intervening quantification.

[1] Professor Burt suggests treating each composite class as a ' determinate ' for a new ' determinable,' and then using the method of maximum probability, if the values are independent, and, if not, the usual weighting formula ($w'R^{-1}$) : see *Factor Analysis of Qualitative Variables*. In theory the standard probability calculus can always be used. Possibly a model could be devised which allowed a ' multivariate multinomial ' distribution to be employed, deduced from its first and second moments.

IV. A NON-PARAMETRIC TEST OF SIGNIFICANCE

All discriminating functions, no matter whether they are based on a mathematical equation or a subjective judgment, can be summarized in a two-way contingency table of observed categories A (in which the individual is observed to fall) against predicted category B. Using the notation of Section III, the columns A_1 to A_p will represent the observed categories. Rows B_1 to B_p will be the predicted classes such that if discrimination were perfect everyone in B_1 would be observed to fall in category A_1, everyone placed in B_2 would fall in A_2, etc. For perfect discrimination the square contingency table would have the diagonal cells filled, but the off-diagonal entries would be zero.

Let the frequency in the ith row and the jth column be denoted by n_{ij}. Let the marginal total of the ith row be B_i and the marginal total of the jth column be A_j. Since B_i is in principle adjustable, we can always, theoretically, make B_i equal to A_i. In this discussion we shall not make that simplification. We shall treat the general case when A_i and B_i are not necessarily equal.

We can now define a pure chance discrimination where correct classifications occur by chance alone. For a pure chance discrimination

$$n_i = \frac{A_i B_j}{N}, \text{ where } N = \sum_{i=1}^{p} \sum_{j=1}^{p} n_{ij}. \qquad (27)$$

Our real interest is in the diagonals of the square contingency table. The misclassifications in the off-diagonal cells may be occurring in a manner contrary to a hypothesis of chance distribution. But this is immaterial except for theoretical analysis of the mistakes. The point is whether the correct classifications are occurring more frequently than chance would allow. A chi-square test would therefore not be exactly appropriate. A contingency table in which the diagonal entries were significantly greater than chance expectation would necessarily have a significant chi-square. But a significant chi-square is not sufficient to show that the discriminating function has made a greater number of correct classifications than would be expected by chance. However, the chi-square test should always be made and if it is not significant the discriminating function on the whole does not do better than chance.

If we are interested in the total number of correct classifications over all observed groups, then we must test the difference between the two quantities. Let

$$o \equiv \sum_{i=1}^{p} n_{ii} \equiv \text{number of persons observed to be correctly classified;}$$

$$e \equiv \sum_{i=1}^{p} \frac{A_i B_i}{N} \equiv \text{theoretical number of correct classifications expected by chance.}$$

Then, the null hypothesis will be that the number of individuals who are correctly classified does not differ significantly from the number who would be correctly classified if the assignment to the categories were based on chance alone: that is, if B_1 individuals were picked at random from the total group of N individuals, the number of such B_1 individuals who will also be in category A_1 will be $\frac{A_1 B_1}{N}$.

In cases where the number of correctly classified individuals, o, is less than or equal to the chance number e, it is not necessary to test for significance. We are only interested in testing those cases where o is greater than e. This is essentially a one-tail test. We are sampling from a population where o can take all values from zero to N. But we are only interested in the portion of the curve from e to N. For small values of N the binomial $\left\{\frac{e}{N} + \frac{N-e}{N}\right\}^N$ should be evaluated. The test of significance would then be to sum all terms of the binomial, from $\frac{N!}{o!\,(N-o)!}\left\{\frac{e}{N}\right\}^o\left\{\frac{N-e}{N}\right\}^{(N-o)}$ to the last term $\left\{\frac{e}{N}\right\}^N$. This sum would give the probability that the number of correct classifications obtained by chance would reach or exceed the number o, on the assumption that e denotes the chance expectation in the general population.[1] When N is large, the binomial distribution is

[1] For a discussion of this procedure in psychological work and references dealing with the mathematical basis, see Burt, *The Backward Child*, App. III, 'The hypergeometric formula from double sample,' where convenient methods of calculation are also described.

approximately symmetric. If the probability distribution is symmetric, the usual t ratio test for the difference between two percentages can be applied, viz., $t = \frac{(o - e)\sqrt{N}}{\sqrt{e(N - e)}}$.

The 99 per cent. confidence interval and the 95 per cent. confidence interval of the binomial distribution have been presented by C. J. Clopper and E. S. Pearson (7). Tables of these values are given by Snedecor (28, pages 4–5).

Even though the observed total number of correct classifications, o, may be significantly greater than chance, it may be suspected that not all predictions are being made with the same accuracy. The n_{ii} for each B_i can be tested against the $e_i \left(\equiv \frac{A_i B_i}{N} \right)$ in the same manner as the overall test of o and e. Here the base number would be the number in the category B_i, rather than N.

The usual t ratio test for the difference between two proportions can be made if it is suspected that category B_i is a better discriminator than B_j. However, note that n_{ii} must be divided by B_i and n_{jj} by B_j to give the correct proportions for the test.

In this way, fairly precise tests may be made of the effectiveness of any discriminating function without having to make any assumptions as to normality of distribution, equal covariance matrices, etc. Johnson (19) gives a practical procedure for determining when the normal curve may be expected to represent the binomial : (1) compute $9N/(N + 9)$; (2) if $e > 9N/(N + 9)$ use the normal curve.

Occasionally we may wish to compare the efficiency of two different discriminating functions on the same data. This leads us to a two by two table. The two rows would represent the individuals (1) who have been correctly classified, and (2) who have been incorrectly classified by the first discriminating function. The two columns would similarly split the groups into correct and incorrect classifications made by the second discriminating function. The four cells would represent (*a*) those individuals correctly classified by both functions ; (*b*) those individuals correctly classified by the first function, but incorrectly by the second ; (*c*) those individuals correctly classified by the first discriminant and correctly by the second ; (*d*) those individuals incorrectly classified by both functions.

With this fourfold table in mind, we may now ask whether p_1, the proportion of individuals correctly classified by the first discriminant, is greater than p_2, the proportion of correct classifications by the second discriminant. For a large enough number of cases, an appropriate test of significance would be

$$\frac{(p_1 - p_2)\,N}{\sqrt{p_1 q_1 + p_2 q_2 - 2r_{12}(p_1 q_1 p_2 q_2)^{\frac{1}{2}}}} = c\,,$$

where N is the total number of individuals to be classified and r_{12} is the point fourfold correlation coefficient. When N is large enough, c will be normally distributed and may be evaluated at the usual levels of significance tabled for the normal curve.

V. SUMMARY

This paper has considered the problems of predicting qualitative attributes from (1) quantitative variables and (2) qualitative variables. The canonical variates which define the minimum subspace within which a set of group means may lie have been derived from the basic equations used in analysis of variance.

The maximum likelihood discriminating functions for the prediction of qualitative attributes from quantitative variables have been described and explained. Explicit formulæ have been given for the case of a multivariate normal distribution with (1) unequal covariance matrices and (2) equal covariance matrices.

A maximum likelihood solution is proposed for the problem of predicting qualitative attributes from qualitative variables.

Some non-parametric tests of significance have been suggested which are applicable to all discriminating functions.

REFERENCES

1. Brown, G. W. (1947). ' Discriminant functions.' *Annals of Math. Stat.*, XVIII, 514.
2. Burt, C. (1921). *Mental and Scholastic Tests.* London : King.
3. Burt, C. (1938). ' The unit hierarchy.' *Psychometrika*, III, 151.
4. Burt, C. (1938). ' The discriminant function ' (unpublished Laboratory Notes).
5. Burt, C. (1944). ' Statistical problems in the evaluation of army tests.' *Psychometrika*, IX, 233.
6. Burt, C. (1948). ' Factor analysis and canonical correlations.' *Brit. J. Psych., Stat. Sect.*, I, 105.
7. Clopper, C. J., and Pearson, E. S. (1934). ' The use of confidence or fiducial limits illustrated in the case of the binomial." *Biometrika*, XXVI, 404.
8. Cochran, W. G., and Bliss, C. I. (1948). ' Discriminant functions with covariance.' *Annals of Math. Stat.*, XIX, 151.
9. Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton : Princeton University Press.
10. Cronbach, L. J. (1949). ' Pattern tabulation.' *Educ. Psychol. Meas.*, IX, 149.
11. Fisher, R. A. (1936). ' Use of multiple measurements in taxonomic problems.' *Ann. Eugen.*, VII, 179.
12. Fisher, R. A. (1938). ' The statistical utilisation of multiple measurements.' *Ann. Eugen.*, VIII, 376.
13. Fisher, R. A. (1939). ' The sampling distribution of some statistics obtained from non-linear equations.' *Ann. Eugen.*, IX, 238.
14. Fisher, R. A. (1940). ' The precision of the discriminant function.' *Ann. Eugen.*, X, 422.
15. Fisher, R. A. (1941). *Statistical Methods for Research Workers.* (8th ed.) Edinburgh : Oliver and Boyd.
16. Guttman, L. (1941). ' An outline of the statistical theory of prediction,' 253–312, in *The Prediction of Personal Adjustment*, edited by Paul Horst. Ann Arbor, Michigan : Edwards Brothers.
17. Hotelling, H. (1931). ' The generalisation of student's ratio.' *Ann. Math. Stat.*, II, 360.
18. Jastak, J. (1949). ' Problems of psychometric scatter analysis.' *Psychol. Bull.*, XLVI, 177.
19. Johnson, P. O. (1949). *Statistical Methods in Research.* New York : Prentice-Hall, Inc.
20. Penrose, L. S. (1947). ' Some notes on discrimination.' *Ann. Eugen.*, XIII, 228.
21. Rabin, A. I. (1941). ' Test score patterns in schizophrenia and non-psychotic states." *J. Psychol.*, XII, 91.
22. Rao, C. R. (1946). ' Tests with discriminant functions in multivariate analysis.' *Sankhyā*, VII, 407.
23. Rao, C. R. (1948a). ' Utilisation of multiple measurements in problems of biological classifications.' *Roy. Stat. Soc. Jour.* (*Series B*), X, 159.
24. Rao, C. R. (1948b). ' Tests of significance in multivariate analysis.' *Biometrika*, XXXV, 58.
25. Rao, C. R., and Slater, P. (1949). ' Multivariate analysis applied to differences between neurotic groups.' *Brit. J. Psych., Stat. Sect.*, II, 17.
26. Rapaport, D. (1945). *Diagnostic Psychological Testing*, Vol. I. Chicago : Year Book Publishers.
27. Smith, C. A. B. (1947). ' Some examples of discrimination.' *Ann. Eugen.*, XIII, 272.
28. Snedecor, G. W. (1946). *Statistical Methods* (4th ed.). Iowa : Iowa State Collegiate Press.
29. Welch, B. L. (1939). ' Note on discriminant functions.' *Biometrika*, XXXI, 218.
30. Wherry, R. J. (1947). ' Multiple bi-serial and multiple point bi-serial correlation.' *Psychometrika*, XII, 189.

APPENDIX

On the Discrimination between Members of Two Groups (Note by C. Burt).[1] Let there be two groups whose means ($\bar{x}_1$ and $\bar{x}_2$), standard deviations (σ_1 and σ_2), and numbers (N_1 and N_2) are known, and whose frequency-distributions are normal but overlap. It is required to find a point on the abscissa, x_i say, which will best serve to discriminate between members of the two groups. To fix our ideas, we may suppose that the first and larger group (denoted by a subscript 1) consists of all the pupils in the ordinary elementary schools, and that the second (denoted by a subscript 2) consists of all those who have been diagnosed as mentally defective by the school medical officers, and transferred to special schools: as tested by the re-standardized Binet scale in 1921, the distribution of the two groups was approximately normal. Our object will then be to find a borderline, in terms of an I.Q., which will minimize the amount of misclassification.

The number of children misclassified will be represented by the sum of two areas, roughly triangular, namely, the upper tail of the left-hand curve added to the lower tail of the right-hand curve. Thus the sum to be minimized will be

$$\int_{x_i}^{+\infty} \frac{N_1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{(x-\bar{x}_1)^2}{2\sigma_1^2}}\, dx + \int_{-\infty}^{x_i} \frac{N_2}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{(x-\bar{x}_2)^2}{2\sigma_2^2}}\, dx = \int_{x_i}^{+\infty} \varphi_1(x)\, dx + \int_{-\infty}^{x_i} \varphi_2(x)\, dx = I_1 + I_2, \text{ say.}$$

To find the value of x which will minimize this sum, we must differentiate it with respect to x, and set the first derivative equal to zero. The first derivative will be

$$\frac{dI_1}{dx} + \frac{dI_2}{dx}.$$

By the usual rules for differentiating a definite integral with respect to a lower or an upper limit, this expression $= -\varphi_1(x_i) + \varphi_2(x_i)$. On setting this equal to zero, we obtain $\varphi_1(x_i) = \varphi_2(x_i)$:

that is,
$$\frac{N_1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{(x_i-\bar{x}_1)^2}{2\sigma_1^2}} = \frac{N_2}{\sigma_2\sqrt{2\pi}}\, e^{-\frac{(x_i-\bar{x}_2)^2}{2\sigma_2^2}}.$$

But these expressions also represent the ordinates of the curves 1 and 2 at the point x_i. We have thus shown that the required value for x_i is the value that marks the point on the base-line between the two means, *at which the ordinate has the same height for both curves;* and this ordinate in turn marks the point where the two curves intersect. Hence, as might have been guessed on intuitive grounds, the best way of discriminating between the two distributions is to cut between the two overlapping curves at the point where the trough between them is lowest.

To express x_i as an explicit numerical function of the constants describing the two curves, we have merely to take logarithms and solve for x_i. We thus obtain the formula given in the text (p. 95 above, and *Mental and Scholastic Tests*, 1921, p. 165). It was on this basis that an I.Q. of 70 was suggested as the borderline for the mentally defective. It should be noted that the proof assumes that we already know the proportion of N_1 to N_2; and the known values must be used, not—as is so often done—either equal numbers or the arbitrary numbers contained in the samples selected by testing. Since the two triangular tails are not exactly equal, the application of the formula will *not* in practice yield *exactly* the required numbers N_1 and N_2. If the psychologist's instructions require him to fill exactly N_2 vacant places in the special schools, then another procedure would have to be used, viz., selecting the dullest N_1 on the I.Q. scale (see *loc. cit.*, p. 167). Further, if his business is to find the best-weighted battery of tests to discriminate two groups, classified on the basis of what is really a continuous, quantitative, and approximately normal variable (such as 'intelligence' or 'neuroticism'), then a discriminant function derived from a criterion that is qualitative only in form is likely to lose much of the information that might have been gained by using a graded criterion.

[1] The following extract from (4) has been appended at Mr. Lubin's request.

THE INFLUENCE OF DIFFERENTIAL WEIGHTING

By CYRIL BURT

I. *The Problem of Weighting.* II. *The Effects of Different Types of Weighting.*
III. *Weighted Sums as Factors.* IV. *Experimental Studies of Differential Weighting.*
V. *Summary and Conclusions.*

I. THE PROBLEM OF WEIGHTING

Weighting in Psychology. In considering how to combine the marks obtained from an examination that includes several question-papers, or from a test battery comprising a number of subtests, one of the most puzzling problems is that of differential weighting. Is anything to be gained by assigning different weights to the component tests or papers, and if so what are the best methods to adopt? In spite of their apparent importance, surprisingly little attention has been given to the issues involved. What Freeman wrote in 1926 is even truer to-day: "Weighting has come to be far less commonly used than it was a number of years ago" [1]; and most practical psychologists who comment on the matter are content to quote Guilford's statement that "weighting is not worth the trouble"—a statement, I fancy, which was intended to apply not so much to the tests in a battery as to the items in a test.[2]

In their book on *The Scientific Study of Educational Problems*—one of the few later publications that touch upon the wider problem—Monroe and Engelhart have briefly reviewed the relevant literature. They point out that the majority of American writers have been concerned chiefly with the influence of weighting on *reliability*, and that the conclusion generally drawn is that "weighting does not affect the reliability to any marked degree." [3] On the other hand, as they rightly say, "the effect of weighting on the *validity* of the measures has received relatively little attention" (11, p. 190). Accordingly, before turning to the statistical questions that appear to be involved, it may perhaps help those who are more interested in practical methods than in technical discussions if I begin by describing in some detail the situations in which such problems have come to the fore.

(a) *Practical Importance.* The earliest instances are to be found in the field of ordinary examinations. For example, in the elementary school system teachers were accustomed to promote pupils year by year on the basis of a composite mark obtained in terminal examinations in all the

[1] *Mental Tests*, 1926, p. 275. In support of his own conclusions he quotes Douglass, H. R., and Spencer, P. L., 'Is it Necessary to Weight Exercises in Standard Tests?', *J. Educ. Psych.*, XIV, 1923, pp. 109–12. For later discussions, coming to much the same views, see Corey, S. M., 'The Effect of Weighting Exercises in a New Type of Examination,' *J. Educ. Psych.*, XXI, 1930, pp. 383–5, Potthoff, E. F., and Barrett, N. E., 'A Comparison of Marks, Based on Weighted and Unweighted Items,' *ibid.*, XXIII, 1932, pp. 92–8, Scates, D. E., and Noffsinger, F. R., 'Factors which Determine the Effectiveness of Weighting,' *ibid.*, XXIV, 1933, pp. 280–5.

[2] *Psychometric Methods*, 1936, p. 448.

[3] This no doubt arises from the facts that, in the theory of least squares, weighting is discussed chiefly in connection with the allowances to be made for differences in reliability (2, 9), and that, in the theory of psychophysical measurement, it is again the differences in the reliability of the observations that require some form of weighting to be introduced (cf. Guilford, *loc. cit.*, pp. 54, 176, and refs.). But, as I pointed out in the memorandum cited below, the problems with which the practical psychologist is concerned depend far more on validity, and thus resemble those arising in the computation of composite 'index numbers' in economics (a point also made by Edgeworth: cf. 1 and 8).

subjects of the curriculum ; certain pupils were not infrequently recommended for a backward class or for possible certification as mentally defective on the ground of low marks ; other pupils were awarded scholarships on the results of a composite examination in a prescribed set of subjects. Discussions continually arose as to the relative weight to be attached to the different subjects : should not Arithmetic be given a higher weight than English, and should not both receive a far higher weight than Manual work ? In such cases, it is to be noted, the interest centres on the single composite mark, intended to represent the ' highest common denominator ' or ' factor ' ; and this factor here relates to a more or less closed and restricted field—namely, the fixed group of subjects forming the elementary syllabus.

Similar questions arose for discussion during the investigations of the *International Institute Examinations Enquiry*. In a memorandum which I was asked to prepare, an attempt was made to derive certain general formulæ which would indicate the effects of equal and of differential weighting on reliability and validity ; a factorial procedure was suggested ; and arithmetical illustrations were given comparing the results obtained by different methods.[1] The main conclusion emerging was that the chief danger lay, not so much in the neglect of carefully calculated differential weights, but in the unrealized presence of random or unintentional weighting—particularly that due to accidental differences in the standard deviations of the examiners' marks. In the published report, to save space and excessive technicality, the formulæ and their algebraic derivations were not reproduced in full ; and both the inferences drawn and the practical recommendations made have since elicited some criticism. More recently, in reporting an investigation carried out for the Scottish Council for Research in Education, and after fully discussing ' The Weighting of Measures,' Dr. McClelland and his colleagues conclude by saying that, " in view of our results, we find it impossible to lay down a single principle, even of the most general character " (14, p. 83). And finally the whole problem has cropped up afresh in connexion with the weighting of tests used in the fighting services during the war and the weighting of examination marks and other assessments for the allocation of pupils to different types of secondary schools.

(b) *Theoretical Importance*. From a theoretical standpoint the question has a still wider interest. Without a clear appreciation of the effects of differential weighting, the nature of the various modes of factor analysis can hardly be understood. But in theoretical discussions on factor analysis the factors envisaged are usually taken as defined, not by denotation (as something common to a definite and circumscribed curriculum of subjects or performances), but rather by connotation (as a quality distinctive of a certain field or domain, of which the processes actually tested or measured form merely specimens or representative samples). Nevertheless, from a statistical standpoint, " a factor is still nothing but a weighted average of a set of measurements empirically obtained, being weighted according to the degree in which it demonstrably represents the field in question "[2] ; and, if this description were kept in mind, the common confusion between factors as statistical components and factors as concrete entities (such as ' abilities,' ' propensities,' and the like)—a confusion against which Professor Thomson has so rightly and so often protested—would less frequently occur.

There is a further confusion which has not so frequently been noted, namely, between the different types of statistical component, regardless of any psychological interpretation. Thus, any factor which has positive saturations for all examples belonging to the same broad category, field, or genus is constantly described as a ' general factor,' regardless of the particular weighting procedure by which the so-called general factor has been extracted : in the field of cognitive processes, for instance, the ' highest common factor ' in a bipolar analysis, the ' basic factor ' in a group factor analysis, the first ' principal component ' in a Pearson analysis, and ' Spearman's g ' are all treated

[1] Burt, C., ' The Analysis of Examination Marks,' *ap.* Hartog, P., and Rhodes, E. C., *The Marks of Examiners*, 1936, p. 303, cf. pp. 245f. As will be seen from the memoranda, Dr. Rhodes was at first highly doubtful, on quite general grounds, about the validity of factorial procedures, and proposed instead to calculate the ' ideal mark ' by " taking a weighted average, the weights being proportional to the degrees of precision, which are inversely proportional to the squares of the standard deviations " (p. 193). Subsequently, however, he reached very much the same formulæ as myself (cf. p. 319) ; and in a later publication of great interest showed how factorial methods might usefully be applied in at least one non-psychological field (Rhodes, E. C., ' The Construction of an Index of Business Activity,' *J. Roy. Stat. Soc.*, C, 1937, pp. 18f.).

[2] Cf. Burt, C., *Factors of the Mind*, p. 74. Spearman, it is true, regularly objected that, as he interpreted the term, a factor was not, and could not be, either an average or a weighted sum (cf. *Abilities of Man*, pp. 61–5). I have, however, already attempted to meet his objections (*loc. cit.*, pp. 211f.), and need not discuss them further here. I may add that, when I first introduced such phrases as ' highest common factor ' (3, 4) and suggested ways of calculating it, I had in mind a ' factor ' in an abstract mathematical sense. However, now that the term is used so frequently to designate a concrete quality, it might be better (as I have suggested elsewhere) to adopt the more accurate term ' component ' for the factor *before* identification, and reserve the term ' factor ' for the ability or other concrete quality with which that component is eventually identified.

as if they referred to precisely the same fundamental concept, even by those who would scrupulously refrain from identifying that concept with ' general intelligence.' [1]

Although in this paper I shall treat the numerical constants that we are trying to determine as no more than weighting coefficients in an equation of estimation, I do not mean to deny that some of them may conceivably have a theoretical significance of their own. Just as the factor-measurements, if they could be appropriately selected and accurately calculated, would prove to be important constants for describing a given individual, so, I believe, the factor-saturations obtained with a suitable choice of tests would often turn out to be important constants describing certain fundamental types of mental activity. In quantitative psychology, the enquiries which lead to a factor analysis mark only the initial stages of our work. Having by that means satisfactorily classified and defined the chief modes of independent variation, we must next press on to determine, wherever we can, those functional equations which are required for laying the foundations of a sound theory of experimental psychodynamics. In such a theory what we now call factor-saturations will, I imagine, play a new role, similar to that of specific heats and dielectric constants in physics or the valencies of the elements in chemistry. This seems confirmed by several recent studies on mental work. But I here refer to it solely to remind the reader that I am far from regarding factor analysis and the calculation of weighting coefficients as the final goal of quantitative research.

Partial Regression Equations. In what follows, we shall be concerned merely with the problem of combining measurements, which are assumed from the outset to form acceptable (though doubtless somewhat fallible) indications of some specified dispositional property—some generalized characteristic which we desire to measure for practical purposes, and which for convenience can be called a factor. Whether such a factor exists and what is its nature are questions we may suppose to have been already decided. The school psychologist, for example, may have to accept the fact that within the schools concerned there is only one line of promotion, and consequently that the promotion of this pupil or of that will depend on their marks for what I called a factor of ' general educational ability ' [2] ; and his immediate problem will be simply to decide what are the relative weights to be attached to papers or tests in English, Arithmetic, Reading, Spelling, Handwork, and the like.

In the earliest factorial studies the main object was to procure individual measurements for the broader factor assumed to underlie all forms of cognition—' general intelligence,' as it was commonly termed. As soon as it was realized that a quality like ' intelligence ' could best be estimated, not by applying a single test (as Spearman and others originally assumed), but by combining a wide variety of tests (as Galton, Binet, and their followers had maintained), the question at once arose as to whether these different tests should be differently weighted, in accordance with their difficulty, their nature, their reliability, or their validity, and, if so, how the weighting coefficients should be computed. Thus, Yerkes, in his ' Point Scale ' revision of the Binet series, contended that we ought to prescribe different maxima for the different test-problems, and put forward various *a priori* considerations for their determination.[3] My own contention was that, in principle, " each test should be weighted, not equally, nor yet arbitrarily, but according to an empirical *regression coefficient* based on its ascertained *correlation* with intelligence itself." [4]

[1] This type of confusion is analogous to that which old-fashioned logicians used to point out between a ' generic *differentia* ' and a ' generic *proprium* ' (a *proprium* is a ' property ' in the last of the four traditional senses recognized by the scholastic writers, namely, *id quod pertinet omni et soli et semper*). Thus, to take the stock scholastic instance, if we define man as a ' rational animal,' rationality is his ' *highest* common factor ' ; ' ability to laugh ' is merely *a* common or ' general ' factor, not the most essential or fundamental.

[2] He is, of course, often tempted to ask how far such a promotion scheme can really be justified ; and in that case one further advantage of the statistical treatment suggested is that it enables us to offer a definite reply to this question also. Thus, up to the age of 10, the single factor of *general* educational ability appears to account for nearly 50 per cent. of the individual differences in the achievements of the various subjects of the curriculum ; whereas each of the remaining factors (*special* educational abilities) accounts for only 10 per cent. or less. On the other hand, at the age of 12 the general factor accounts for little over 25 per cent. : (cf. 6, p. 266 ; also *Brit. J. Educ. Psych.*, IX, p. 55, XIII, p. 132, and refs.).

[3] Yerkes, R. M., *et al.*, *A Point Scale for Measuring Mental Ability*, 1915. For criticisms of the weighting proposed, see *Mental and Scholastic Tests*, p. 72. Cf. also Binet, A., ' On the Necessity of Establishing a Scientific Diagnosis of Inferior Intelligence,' *L'Année Psych.*, XI, 1905, pp. 163f.

[4] Burt, C., *Mental and Scholastic Tests*, 1921, p. 74. The first suggestion that regression coefficients should be used in this way was contained in my article on ' The Measurement of Intelligence by the Binet-Simon Tests,' *Eug. Rev.*, VI, 1914, pp. 140–52.

In support of this suggestion, I argued that " if we adopt Professor Pearson's device of ' multiple correlation,' it can easily be demonstrated that a far better assessment can be achieved with a representative sample of tests, covering a number of different cognitive processes, each appropriately weighted, than can be obtained with only a single test, or even a mere total of marks from several tests taken as they stand." In seeking in this way to secure more trustworthy measures for the ' general factor ' of intelligence, we are in effect taking samples or specimens of different types of intelligent performance, much as, in endeavouring to procure a trustworthy estimate of (say) the proportional amount of water in human muscle, the physiologist examines specimens of all the chief types of muscular tissue. The only novelty is that we attempt to weight each specimen according to its representative value.

External and Internal Criteria. By what criterion, however, are we to determine the correlation of each test with the quality it is intended to measure ? In the case of general intelligence, it was at first assumed that the independent estimates supplied by teachers or other observers would provide a valid standard. On that assumption the diagnostic value of each specimen test could be assessed by comparing or correlating its results with this external criterion, and appropriate weights could then be computed by the principle of least squares, in such a way that the resultant mark should yield the closest approximation to the estimates which the criterion supplied. Unfortunately, in many investigations no such external standard was available, and in others the independent estimates provided were not sufficiently trustworthy to be used for that purpose. The method of factor analysis was accordingly developed, specifically to meet cases of this kind. But essentially the same formulæ still seemed applicable for calculating the weights, although the criterion was now based on internal rather than on external evidence.

Substitutes for Regression Equations. The full procedure, however, though plausible enough in theory, proved decidedly laborious in practice. It entailed lengthy calculations, first to determine the regressions or weights and then to compute the weighted sums. Accordingly, for practical purposes various simplified methods were commonly substituted, without any very full justification being offered for their use.

(i) *Unweighted Sum of Standardized Measurements.* In the early days of factorial work, the most usual procedure was to reduce all the test-measurements to a comparable unit, either by using standard measure, or by converting the test-measurements into ranks, and then to sum or average the measurements so standardized. Tests which had a low correlation with all the other tests, and therefore with the general factor, were generally rejected. Thus, in my 1909 article on ' Experimental Tests of General Intelligence ' (3), out of the whole series of twelve tests, I retained the best six, and averaged the ranks for those six. This was found to raise the correlation from 0·76 (the coefficient furnished by the best single test) to 0·85.

(ii) *Saturations as Weights.* In later work, where more accurate factor-measurements were required (e.g., for the supplementary factors), the saturation coefficients themselves were not infrequently used as approximate weights. But sometimes a test with a positive saturation may have a negative regression (see Appendix, p. 124).

(iii) *Simplified Weights.* More often a rough and ready compromise between the two procedures was adopted. Instead of saturation coefficients, which entailed multiplication by two- or three-figure decimal fractions, simple unit-figure weights (e.g., 1, 2, 3) were adopted for the better tests, and the poorer tests omitted as before (a more recent example of this procedure will be found in the Scottish investigation on selection for secondary education, 14, pp. 82f.). This kind of weighting can be most easily introduced by allowing 1, 2, or 3 marks for each correct answer according to the weight to be given to the several tests in a carefully selected battery, or by prescribing as the maximum mark for certain tests or question papers a figure two or three times as large as that for the others. In compiling booklets of tests, a common plan was to increase or reduce the number of problems in each test in such a way that the marks obtained for each test at once eliminated the effects of a varying standard deviation, and gave the better tests a higher loading, without any explicit multiplication being required when computing the totals.

Whatever method of weighting is adopted, it should be remembered that the real effects of a set of weights must be assessed in terms of the *relative contribution of each weighted component to the total variance ;* and this will depend, not only (i) on the assigned or ostensible weights, but also (ii) on the standard deviation of each component, and (iii) on its correlation with each of the other components (see Appendix, p. 124).

Weighting in a Non-Psychological Research. It is interesting to note that, in attempting to determine the relative ' crop productivity ' of different British counties, Professor Kendall (13) has recently used two alternative devices, not unlike those I have described. They are, as he says, " based on psychological work, and measure a *factor* which may reasonably be regarded as covariant with productivity by regarding crop yields as analogous to test scores." His first method is to calculate what I should call ' factor-measurements ' for the ' highest common factor,' i.e., for the ' general factor of productivity ' ; and for this purpose he employs a partial regression equation derived in accordance with Pearson's method of ' principal axes.' His second method is a simplified procedure similar to the first of those mentioned above, namely, simply ranking the counties for their production in each of the main types of agricultural produce, and then taking the average rank to indicate general productivity. He finds that " the closeness of the results, which are reached by widely different methods, is surprising " : the correlations between the two sets of figures, obtained for four different years, all fall between 0·95 and 0·97 (inclusive). Thus, so far as these examples go, " the average rank tends to give the same ranking as one type of general factor loading." He remarks that he has " been unable to find any mathematical proof that a close correspondence will always exist " ; and he asks " whether a similar problem has been studied in psychology " (13, pp. 25, 45).

II. THE EFFECTS OF DIFFERENT TYPES OF WEIGHTING

Special Problems. In examining the general effects of different modes of weighting, the problems involved can, I fancy, be most usefully studied by considering the process in four stages, namely, (*a*) *a priori* or subjective weighting, (*b*) random weighting, (*c*) equal weighting, and (*d*) *a posteriori* or *ad hoc* weighting (i.e., weighting based on empirical data analysed specially for the purposes of the particular research).

(*a*) A Priori *Weighting.* In the early days of mental and scholastic measurement,[1] the methods of weighting proposed were usually based on subjective considerations. The highest weights were assigned to those aspects of the subject which were deemed, on *a priori* grounds, to be of greatest moment, or to those measurements which were presumed to be the most reliable. Such principles are still often followed in the field of education. Examiners, for instance, frequently allot different maxima to different question-papers according to the supposed importance of the subjects ; and, as a rule, no very clear distinction is drawn between ' weighting for value ' and ' weighting for validity.'[2] *A priori* weighting also seems needed where the criterion appears to be composite in content, i.e., to be influenced by several components, which may be independent, alternative, or in some cases even compensatory (as in occupational success).

[1] The earliest discussion of the need for a systematic determination of weights in attempts at mental measurement is to be found in Edgeworth's papers on ' The Statistics of Examinations ' and kindred topics : *J. Roy. Stat. Soc.*, LI, 1888, pp. 599–635 ; LIII, 1890, pp. 460–75, 644–63. Cf. also Galton, *Brit. Ass. Ann. Rep.*, 1889, pp. 471f.

[2] I here use the term ' validity ' to include ' precision ' or ' reliability ' on the ground that diminished reliability is one cause of diminished validity. By ' weighting for value ' I mean attempts in which the weights proposed are based on judgments of practical utility or cultural worth rather than on diagnostic validity or reliability. Thus, at a meeting of examiners for the junior county scholarship examination, certain members urged that the maximum allotted to the English paper should be higher than that allotted to the Arithmetic paper, " because in a civilized English-speaking community it is more important for its members to write English than to do sums." The validity of each paper as an index of innate general ability was not discussed. For an interesting suggestion as to how " weighting for social importance " might be carried out, see Kelley, T. L., *Essential Traits of Mental Life*, 1935, pp. 70, 92–5 : especially equations (iii) and (47).

In this respect the mode of weighting was not unlike that employed in calculating composite 'index numbers' [1] by early economists ; and thus open to very similar criticisms (cf. 7, p. 212, for a remark which applies admirably to the psychologist's computations). *A priori* weighting is occasionally used in other scientific fields. But, according to those most competent to give a judgment on the matter, "'empirical' weighting, that is, weighting according to individual judgment" (or, as the psychologist would prefer to term it, 'subjective' or 'impressionistic' weighting), "is always a matter of difficulty, and at best its results must remain open to objection." [2] Accordingly, I venture to suggest that even here a factorial procedure might prove suggestive.

Even where those concerned are unanimously agreed that trait x (say) has a higher social or cultural value than trait y, and therefore deserves a weight which is n times as large as that assigned to y ($n > 1$), there seems to be no generally recognized principle for deciding what numerical value shall be assigned to n. Accordingly, I suggest what may be called the principle of virtual frequency. To allot to x a weight which is n times that allotted to y, means that we are in effect treating each of the x-measurements as if they had been based on *n times as many independent observations* as the y-measurements.

I may add, however, that, alike in psychological testing and in school or university examinations, wherever I have been able to check the results of such subjective weighting against objective criteria, its value has proved extremely doubtful. I should thus readily endorse the view expressed by Sir Percy Nunn. In a memorandum on 'Methods of Marking' (dealing primarily with the Teachers' Certificate Examination), he has stated that, "with rare exceptions, no satisfactory case can be made for allotting different weights to different examination papers merely on personal or *a priori* grounds. The wide differences between examiners as to what the weights should be, and the erratic effects on mark-lists, suffice to demonstrate that in most cases subjective weighting is just random weighting. What the precise effect of such random weighting may be it is scarcely possible to predict. But one thing can be said with certainty : its general tendency must be to reduce, seldom to enhance, the accuracy of the final marks." [3]

(*b*) *Random Weighting*. As I have already noted, it is too often forgotten that, when a set of measurements or marks are summed or averaged just as they stand, a certain amount of weighting is unintentionally, and often quite unconsciously, introduced, owing to the fact that the raw marks or measurements have different standard deviations. Such differences in the scatter are likely to have no relation either to the value of the separate tests, or to any common factor underlying the whole set of tests or to any valid criterion which the tests or marks are intended to estimate. Thus their effects are equivalent to that of random weighting. Since, then, both deliberate weighting and unintentional weighting so often tend to introduce a large element of random weighting, it seems desirable to determine, if we possibly can, what (in Nunn's words) "the precise effect of such random weighting" is likely to be.

Let us, therefore, begin by examining the effect of weights, selected absolutely at haphazard, on the weighted sums so obtained. Let us suppose that we have n sets of assessments, $x_1, x_2, \ldots x_n$, each set measuring the same attribute, x, with different degrees of accuracy, or, it may be, measuring some different but typical manifestation of the same attribute. (If the reader wishes to keep a concrete instance in mind, he may think of the assessments as the result of applying n tests of intelligence to N school children.) Our ultimate aim is to combine these n measures into a single composite measure by means of some appropriate linear function. Let the set of weights for one such linear function be designated U_i, and for

[1] It was for this reason that I ventured to suggest that the methods of factor analysis might be applied to this and other problems in economics (10, p. 313). Bowley's definition of index numbers (5, p. 196) would serve admirably as a definition of factors. The chief difference (which curiously enough is not brought out explicitly by his definition) arises from the fact that index numbers are commonly used to measure variations which are related to time, and hence are expressed as percentages of an arbitrary standard instead of as deviations about the mean. I may add that index numbers, like factors, are of two distinguishable types : (*a*) they may refer to a domain defined connotatively (as with Jevons' monetary index) ; (*b*) they may refer to a limited domain defined denotatively (as with the index for cost of living) : (cf. p. 106 above).

[2] Freund, I., *The Study of Chemical Composition*, 1904, p. 85. Cf. Kohlrausch, *Physical Measurements*, 1894, especially chapter on 'Errors of Observation,' and Ostwald, *Physico-Chemical Measurements*, 1902, chap. I, 'Calculation.'

[3] This opinion was repeated in a later memorandum submitted by Sir Percy Nunn to the International Institute Examinations enquiry.

another W_i ($i = 1, \ldots n$). Then the two alternative composite measures which we propose to compare may be written

$$y = U_1x_1 + U_2x_2 + \ldots + U_nx_n , \qquad \text{(i)}$$

$$z = W_1x_1 + W_2x_2 + \ldots + W_nx_n . \qquad \text{(ii)}$$

We shall assume that x_i, y, and z are each to be in unitary standard measure.

Now let $\bar{r}$ denote the average of the $\frac{1}{2}n(n-1)$ inter-correlations, $\overline{U}$ and $\overline{W}$ the means, and σ_u^2 and σ_w^2 the variances of the weights. It is then easy to show (see Appendix) that, if the weights have been chosen at random, the correlation between the two composite measures will be approximately

$$r_{yz} = 1 - \frac{(1-\bar{r})}{n\bar{r}} \cdot \tfrac{1}{2}\left(\frac{\sigma_u^2}{\overline{U}^2} + \frac{\sigma_w^2}{\overline{W}^2}\right), \qquad \text{(iii)}$$

$$\text{or } 1 - \frac{(1-\bar{r})}{n\bar{r}} \cdot \frac{\sigma_w^2}{\overline{W}^2} , \qquad \text{(iv)}$$

when the means and standard deviations of the weights are the same.

These two expressions indicate very clearly what will be the general effect of random weighting in various actual cases. The following conclusions may be drawn.

(1) The larger the number of tests employed (i.e., the higher the value of n), the higher will be the correlation between the two weighted sums. As the number of tests is increased indefinitely, so the resulting correlation, r_{yz}, tends to become perfect, even though the two sets of weights have been chosen absolutely at random.

(2) The larger the average inter-correlation between the several tests (i.e., the higher the value of $\bar{r}$), the higher will be the correlation between the two weighted sums. As the average inter-correlation approaches 1·00, so the resulting correlation, r_{yz}, tends to become perfect, even though the weights have been chosen absolutely at random. Conversely, the lower the average inter-correlations, the lower will be the value of r_{yz}. An important corollary follows. If, in order to be sure of assessing a factor from different angles, we have employed a highly heterogeneous set of tests, which have low inter-correlations, and if for practical reasons we can only include a small number of tests in our battery, then the choice of appropriate sets of differential weighting becomes of special importance.

(3) The smaller the standard deviations of the weights, the higher will be the resulting correlation between the two weighted sums. If in both sets the standard deviation is zero (i.e., if within each set the weights are identical, though the weights in one set may not be identical with those in the other), then the correlation will be unity. Conversely, with random weights, the larger the standard deviations, the lower will be the correlation. This confirms the familiar observation that, provided the individual weights do not vary widely, the agreement between the two weighted sums tends to be pretty close, whereas when the range of the weights varies widely, the agreement may be greatly reduced.

(4) If in both sets the mean weight is large in comparison with the standard deviation, then once again the agreement between the two weighted sums will be close. It is thus the relative size of the differences between the weights, not their absolute size, that affects the resulting correlation.

(5) Finally, for the correlation between the weighted sums to be positive, both the mean weights must have the same sign. In particular, if in each set the vast majority of the weights or all the positive weights are positive throughout, then the resulting correlation r_{yz} will be positive. This has a special bearing on the calculation of measurements for the ' first factor ' in factorial procedures. Whether it is a ' general ' factor, a ' basic ' factor, or a first ' principal component,' it is in most psychological enquiries likely to have positive saturations for all the tests or traits in the battery (with the possible exception of one or two variables which may act as ' suppressors '). Hence, unless the number of tests or traits is small, gross inaccuracies in the weighting coefficients will impair the resulting measurements far less than might be supposed.

(*c*) *Equal Weighting*. Let us now turn to the case in which the n standardized tests are given equal weights. This is of special interest on both practical and theoretical grounds. From a practical standpoint it provides by far the simplest procedure for combining a set of measurements into one composite measurement ; from a theoretical it is the method of combination implicitly adopted in the analysis of variance (see 15, Table III).

Let us consider first the common situation in which the n tests have been administered twice to the same group. For example, in the 1909 enquiry the six tests mentioned above were applied first

by myself and secondly by Dr. Flugel. A straightforward deduction from Pearson's product-moment formula yields the following result :[1]

$$r_{\bar{x}\bar{x}'} = \frac{\Sigma(x_1 + x_2 + \ldots + x_n)\,(x_1' + x_2' + \ldots + x_n')}{\sqrt{\Sigma(x_1 + x_2 + \ldots + x_n)^2}\ \sqrt{\Sigma(x_1' + x_2' + \ldots + x_n')^2}}$$

$$= \frac{\Sigma r_{ii'} + \Sigma\Sigma r_{ij}}{n + \Sigma\Sigma r_{ij}} = \frac{\bar{r}_{ii'} + (n-1)\bar{r}_{ij}}{1 + (n-1)\bar{r}_{ij}} \qquad \text{(v)}$$

where the dash denotes the second application of the test, and $r_{ii'}$ the reliability coefficients (it is assumed that r_{ij} and $r_{i'j'}$ have been pooled to determine what is here called r_{ij} ; $i \neq j$). In the example cited the formula yields $r_{\bar{x}\bar{x}'} = \frac{0{\cdot}682 + 5 \times 0{\cdot}423}{1{\cdot}000 + 5 \times 0{\cdot}423} = 0{\cdot}898$. On calculating the correlation direct from the averaged ranks, the value reached was 0·902. (The slight difference is due to rounding.)

If we know the correlation of each test with the criterion (c, say), then the correlation of the unweighted sum with the criterion will be

$$r_{\bar{x}c} = \frac{\Sigma r_{ic}}{\sqrt{\Sigma\Sigma r_{ij}}} \qquad (i, j = 1, \ldots n)$$

$$= \frac{\bar{r}_{ic}\sqrt{n}}{\sqrt{1 + \bar{r}_{ij}(n-1)}} \qquad (i \neq j)\ . \quad \text{(vi)}$$

By solving for n in one or other of these equations it is possible to estimate roughly how many subtests will be required to yield a correlation of a required amount.[2] If the criterion is an internal criterion, based on simple summation, the expression thus reached simplifies to the square of the expression reached in equation (v), with $r_{ii'}$ now representing communalities rather than reliabilities. The reader can easily deduce for himself the corresponding formulæ for an internal criterion derived by weighted summation or any other procedure.

The same argument shows that the correlation of each test (say, the kth) with the average of all the tests is

$$r_{kg} = \frac{\Sigma r_{kj}}{\sqrt{\Sigma\Sigma r_{ij}}} \qquad (i, j = 1, \ldots n)\ , \quad \text{(vii)}$$

where r_{ii}, the self-correlation, is here taken to be 1·00. Now, with the majority of correlation tables obtained in psychological testing, an application of this formula would at once reveal that the tests employed have a widely different correlation with the ' general factor,' estimated provisionally as the unweighted average of all the tests. And this naturally suggests that it is a mistake to give all the tests precisely the same weight. Hence we are led to enquire how the best set of differential weights can be procured.

(*d*) *The Best Weights.* If we ignore the possible existence of specific factors, the most obvious method would be to seek the ' line of closest fit.' This was the procedure proposed by Karl Pearson, who defined ' closest ' in terms of least squares.[3] He further pointed out that the requisite weights could be conveniently obtained by seeking the " largest principal axis of the correlation type ellipsoid." And it is, as Kendall has observed, " a remarkable fact that all the information we need, even to represent the variation in a linear subspace of m dimensions, is contained in the correlation matrix." [4]

[1] The principles from which this and kindred formulæ were derived were first developed by Karl Pearson. To psychologists equation (v) is best known from its application by Brown and Spearman to the problem of increasing the reliability of tests by increasing their length (cf. Brown, W., and Thomson, G., *Essentials of Mental Measurement*, 1925, p. 132 and refs.). Brown points out that his formulæ were " based on demonstrations given in lectures by Professor Pearson " (the source from which many of my own were also derived).

[2] See 10, p. 303, where several numerical illustrations are given. It should be emphasized that, as the assumptions involved are never likely to be precisely fulfilled, the indications must be regarded only as approximate.

[3] Pearson, K., ' On Lines and Planes of Closest Fit,' *Phil. Mag.*, II, 7th ser., 1901, pp. 559f.

[4] Cf.17, § 9. I should myself prefer to say " in the table of covariances and variances." This includes the correlation matrix with unity in the diagonal as a special case. The method, however, can be applied to the unstandardized table of variances and covariances (or square-sums and product-sums) ; and Kelley and others have supposed that the components so obtained would be the same as when obtained from the standardized table of correlations : but it is easy to show that this is not the case. For certain purposes an analysis of the unstandardized covariance matrix seems preferable.

It is instructive to note that the same solution is reached with formulations of the problem that at first sight seem quite different. For example, we may seek (i) weights that minimize the squared discrepancies between the observed *test-measurements* for the individuals and the fit based on the hypothetical factor-measurements, (ii) weights that minimize the squared discrepancies between the observed *correlations* and the fit based on the hypothetical correlations, (iii) weights that maximize the variations between the individuals, i.e., the general *factor variance*, or (iv) weights that maximize the square-sum of the *multiple correlations* between the observed factor-measurements and the hypothetical factor-measurements : all lead to Pearson's method of ' principal components.'

Kendall has stated that " this approach is the first which the statistician would consider, and he [the statistician] may be a little surprised to find that the psychologists do not give it pride of place." But in the early days of factorial work—say, from 1909 to 1924—all those who attempted to factorize complete correlation tables would, I think, have given Pearson's method, if not pride of place, at least priority of place. The reasons for the modifications I have explained elsewhere (16).

The chief objection to the method was, of course, the difficulty of carrying out the calculations ; and this, as we have seen, led to the substitution of simpler schemes of weighting and in particular to the use of equal weighting. But, if we may judge from recent discussions of factor analysis by statisticians, the change which has aroused the most widespread criticism is the use of reduced self-correlations. My reasons for making the reduction [1] have been given elsewhere, and may be briefly summarized as follows. With self-correlations of unity we are led (i) to explain the observed correlations by more factors than the data warrant, and (ii) to represent the specific factors by obviously artificial bipolars. (iii) Further, Pearson's procedure, which seeks the ' line of closest fit ' by the method of least squares, places all the test-results on the same footing. Now in psychological work the measurements supplied by tests of different kinds are apt to differ widely in their precision and their validity. For example, among the twelve intelligence tests I employed in my first investigation (3), the best had correlations with the criterion amounting to 0·80 or thereabouts ; the poorest had correlations of only 0·20 or less. But it is surely unreasonable to insist that the calculated estimates of intelligence should fit the measurements furnished by the poorest tests as closely as they fit the measurements furnished by the best. An allowance for such differences can be made by what I termed ' correction for specificity ' ; and this in turn entails substituting reduced self-correlations for self-correlations of unity : (see 16, p. 111 and refs.).

III. WEIGHTED SUMS AS FACTORS

A General Method of Factorization. So far our discussion has been restricted to the type of weighting problem most commonly encountered in actual practice, namely, that in which we assume a single dominant factor common to all our tests or question-papers, and are concerned solely to determine a final set of marks measuring this common factor as accurately as possible. Factor analysis, however, goes further than this : it reveals the existence of other factors, assesses their relative importance, and claims to supply a method of estimating measurements for these additional factors, whenever they are required. For the supplementary factors the method of simply summing the scores or their residuals may still be used. But, as can readily be demonstrated, it is likely to be far more inaccurate than for the

[1] From about 1915 onwards this substitution became almost automatic with investigators who calculated saturations from complete tables of correlations. The procedure more recently proposed by Thurstone may at first sight appear the same as my own : indeed, his brief reference to my views might seem to imply that my ' reduced self-correlations ' are identical with his ' communalities ' (*Multiple Factor Analysis*, p. ix). But whereas my reduction is determined primarily by considerations of *statistical significance*, Thurstone's is determined by considerations of *rank*. The ' correct values ' of the ' communalities,' he tells us, " maintain the rank of the side correlations, and this is the minimum possible rank.... If smaller values were taken, the rank would rise " (p. 283). As a result, his values are almost always larger than mine. This will be obvious when we recall that in practice he takes the largest of the observed test-correlations as a convenient approximation to its communality. I cannot see that " the rank of the side correlations " as such can be of any objective importance : with an empirical table it is simply a function of the number of tests we happen to have put into the battery. It is true that the significance of the residuals will in part depend on the number of the tests, but not in the way Thurstone's procedure assumes.

first factor ; and for later factors the loss of accuracy tends to be cumulative.[1] For rough pilot enquiries, no doubt, the method may often be suggestive.

In such cases, however, the question is naturally raised : are we really justified in regarding these rough and ready estimates as representing ' factors ' ? Spearman and those who worked with him insisted that they could not be treated as factors at all ; and this appears to be the general view. But I believe it will not be difficult to prove that such estimations have all the fundamental properties of a factor.

It would, I fancy, now be almost universally admitted that what is commonly called a ' factor ' is essentially a ' weighted sum ' ; yet it is not so widely recognized that any ' weighted sum ' constitutes a ' factor.' To demonstrate this point, however, it will be necessary to define our terms a little more precisely. As Professor Kendall has pointed out, factor analysis has been " developed mainly by men who were not primarily statisticians " ; [2] and in consequence it has " tended to acquire a language of its own." But, although there may be some doubt over minor points, most of the terms employed by psychological factorists can be translated without loss of meaning into the terminology of matrix algebra.

If, then, we agree to regard a table of measurements for n traits obtained from N persons as an $n \times N$ matrix, M_0 say (where $n < N$), we can regard the figures in each row of M_0 as linear functions of r linearly independent $1 \times N$ row-matrices or ' vectors,' chosen to serve as a convenient ' basis ' for all such measurements, where r will in general stand for the rank of M_0 ($r \leqslant n$). Moreover, these selected vectors may or may not be statistically independent. Thus we may write

$$M_0 = FP, \qquad \text{(viii)}$$

where F and P denote what (in the usual psychological context) would be called the ' factor-saturations ' for the traits and the ' factor-measurements ' for the persons respectively. The equation states that the measurement for the jth person in the ith trait can be expressed as a weighted sum, viz.,

$$m_{ij} = f_{i1}p_{1j} + f_{i2}p_{2j} + \ldots + f_{ir}p_{rj} = \sum f_{ik}p_{kj}\ , \qquad \text{(ix)}$$

where f_{ik} are the ' weights ' assigned to the kth factor when reconstructing the ith test. By solving the equations for the p's (or adopting some equivalent procedure) we may also succeed in expressing the p's as linear functions or weighted sums of the m's.

Now if we use f_k (with a single subscript) to designate the kth column in F, and p_k to denote the kth row in P, then we can write $f_k p_k = M_k$, say, where M_k will be an $n \times N$ rectangular matrix of ' rank one,' that is, all its rows will be proportional ; in other words, M_k will form what in factorial jargon is called an asymmetrical ' hierarchy.' It follows from (viii) that M_0 can be analysed into r additive sub-tables of this kind, so that

$$M_0 = M_1 + M_2 + \ldots + M_r = \sum M_k\ . \qquad \text{(x)}$$

If equation (x) is to be fulfilled exactly, the number (r) of unit-rank sub-tables in this series will be identical with the ' rank ' of M_0. By the rank of a matrix is meant the number of linearly independent row-vectors (or column-vectors [3]) contained in that matrix ; and, if an exact factorial

[1] I endeavoured to illustrate this conclusion by comparing the results of factor analysis with those of analysis of variance, since with analysis of variance weights are in effect treated as numerically equal (cf. 15, section on ' analysis by simple averaging,' and Table XII, last two lines). The same point emerges if we turn to the illustrative example provided by Professor Kendall. On using factor analysis to extract supplementary factors, we find that these clearly classify his crops into four groups—cereal, leguminous, roots, and grass. But if we use unweighted figures and try to find a possible classification by working with simple ranks and rank-differences, the grouping is greatly obscured.

[2] Cf. 17, § 1. I would also add " and by writers compelled to use language that would be broadly intelligible to psychologists, teachers, and research-workers, who were not themselves familiar with statistical terminology." Thus a non-mathematical psychologist seems to get some inkling of what factorists mean when they use a phrase like ' highest common factor,' whereas he would feel bewildered if they talked of ' the orthogonal component which makes the maximum contribution to the variance ' ; he understands how to calculate a ' weighted sum,' but, were he an editor, he would in the early days have refused (as Ward once did) to publish a paper which spoke of " such technicalities as linear functions and principal axes."

[3] The number will be the same whether we select rows or columns. From this definition it follows that the rank will give the order of the largest minor determinant in M_0 which is not zero. This latter theorem is usually taken as the definition of rank (Bôcher, M., *Higher Algebra*, 1907, p. 22 ; Aitken, A. C., *Determinants and Matrices*, 1939, p. 60). But for purposes of factor analysis the definition given in the text is more convenient : it is that adopted by Macduffee, *Vectors and Matrices*, 1943, pp. 22, 64.

reconstruction is desired, this will be identical with the total number of ' factors ' precisely obtainable from that matrix.

Linearly Independent Factors. The most general way of effecting such an analysis is to take for the first hierarchical matrix an expression of the type

$$M_1 = \frac{1}{v} s_t s_p , \qquad \text{(xi)}$$

where $s_p = w_p' M_0$, $s_t = M_0 w_t$, and $v = w_p' M_0 w_t$. Here s_p will be a row-vector whose elements are derived by weighting and adding the figures in each column of M_0 ; s_t will be a column-vector whose elements are derived by weighting and adding the figures in each row of M_0 (in the usual context s_p will be a row of sums for persons, and s_t a column of sums for tests) ; v, a scalar (analogous to a variance), will be simply an adjustable divisor introduced to bring the products $s_t s_p$ to terms of the same scale as M_0. So far, it is to be noted, w_p' may be *any row of weights*, and w_t may be *any column of weights*, provided, of course, that the divisor v is not thereby reduced to zero.

Now, an application of the usual criterion for rank will quickly show [1] that, whatever be the rank of the initial matrix M_0, the matrix of first residuals, that is, $M_0 - M_1 = M_{0\cdot1}$, will *necessarily have a rank which is one degree lower.* Moreover, these residuals in their turn may be factorized in precisely the same way. Once again, any set of weights may be used, provided, of course, they are not the same as the first set ; and once again the rank will be reduced. We can repeat the subtraction and factorization r times (r being the rank of M_0), whereupon the last set of residuals will be found to vanish identically.

At the same time we may note that, at every stage, the standardizing divisor may be considered as applied either to the expression $w_p' M_0$ or to the expression $M_0 w_t$ (and their subsequent analogues) or shared between the two (by taking the square root of the divisor). Thus, according to choice, s_p may be regarded as specifying factor-measurements for the persons (standardized if desired), and s_t the weights for the tests or traits ; or alternatively s_t can be regarded as specifying factor-measurements for the traits (standardized if desired), and s_p the weights for the persons. Consequently, the resolution I have described may be effected with a view *either* to analysing and classifying the persons *or* to analysing and classifying the traits. But, whichever happens to be our aim at the moment, the point I wish here to stress is that *any* set of weights—for instance, weights that are equal to one another (say 1, 1, . . . 1), or a weighting based on a single test (say 1, 0, . . . 0), or even a set of weights chosen absolutely at random—can be used to effect an analysis into factors.

Statistically Independent Factors. A more narrowly defined set of factors can be secured if we select our weights so that the factor-measurements will be not merely linearly, but statistically, independent (i.e., uncorrelated). This can be effected by the simple device of taking $w_t = s_p' = M_0' w_p$, and thus avoiding the introduction of a second set of freely chosen weights. We then have $v = w_p' M_0 s_p = w_p' M_0 M_0' w_p$. Equation (xi) now becomes

$$M_1 = \frac{1}{v} s_t s_p = \frac{M_0 M_0' w_p \cdot w_p' M_0}{w_p' M_0 M_0' w_p} = \frac{R_0 w_p \cdot w_p' M_0}{w_p' R_0 w_p} , \qquad \text{(xii)}$$

where R_0 denotes the matrix of observed correlations and M_0 (as usual) is assumed to be in unitary standard measure.

To standardize s_p we must divide it by the root-square-sum. We then have

$$p_1 = v^{-\frac{1}{2}} s_p = \frac{w_p' M_0}{\sqrt{w_p' R_0 w_p}} = w_1' M_0 \text{ (say)} ; \qquad \text{(xiii)}$$

and we can then take

$$f_1 = v^{-\frac{1}{2}} s_t = \frac{M_0 M_0' w_p}{\sqrt{w_p' R_0 w_p}} = M_0 p_1' : \qquad \text{(xiv)}$$

so that f_1 is the column of correlations between the test-measurements (M_0) and the factor-measurements (p_1'). As a corollary we have $f_1 = M_0 M_0' w_1 = R_0 w_1$; and therefore $w_1 = R_0^{-1} f_1$. We now have for the first hierarchical sub-table $M_1 = f_1 p_1$. An incidental advantage of this procedure is that it enables us to calculate the factor-saturations f_1 and the factor-measurements p_1 , direct from the small $n \times n$ matrix of correlations, R_0, instead of from the larger initial $n \times N$ matrix of test-measurements M_0.

We can factorize the residual matrix $M_{0\cdot1} = M_0 - M_1$ in precisely the same way. We then have for the second set of factor-measurements $p_2 = w_2' M_{0\cdot1}$, where w_2' may again be any set of

[1] This statement and the general method of reduction is ultimately due to Lagrange (cf. Bôcher, *loc. cit. sup.*, pp. 131f., and Wedderburn, J. H. M., *Lectures on Matrices*, 1934, pp. 68f.).

weights. And for the second set of factor-saturations we have $f_2 = R_{0\cdot1}\, w_2$. The correlation between the two factor-measurements will now be

$$p_2 p_1' = w_2' (M_0 - f_1 p_1)\, p_1' = w_2' (M_0\, p_1' - f_1) = w_2' (f_1 - f_1) = 0. \qquad \text{(xv)}$$

This result shows that, with *any* two sets of weights, w_1 and w_2, we shall obtain two sets of uncorrelated factor-measurements. (The only restriction is that the divisors $w_1' R_0 w_1$ and $w_2' R_{0\cdot1} w_2$ shall not be zero ; and this evidently prevents us from taking the *same* set of weights for w_2 as for w_1.)

We can continue the process to analyse the successive sets of residuals, until, after deducting the rth hierarchical sub-table, we shall be left with residuals of zero. It follows that we may (if we wish) choose completely random weights for $w_1, w_2, \ldots w_r$, or (if we prefer) unit-weights such as $[\pm1, \pm1, \ldots \pm1]$ (in which case we have to alter the sign-pattern to prevent $w_1 = w_2 = \ldots = w_r$), or again [1, 0, 0, . . . 0], [0, 1, 0, . . . 0], [0, 0, 1, . . . 0], and so on. We can in this way determine the rank of M_0, and resolve M_0 into an additive series of r hierarchical sub-tables.[1]

In practice, of course, with an empirical set of $n \times N$ test-measurements, the rank of M_0 will be n or N, whichever is smaller ; and our interest will be, not to ascertain the complete number of factors, but the *minimum* number of significant factors, that is, what may be called the ' significant rank ' of M_0. Into the further limitations imposed on w_1, w_2, etc., by this or other supplementary conditions it is unnecessary to enter here, since the more specialized procedures then required are already sufficiently well known.

IV. EXPERIMENTAL STUDIES OF DIFFERENTIAL WEIGHTING

An Attempt at Empirical Verification. To illustrate the effects of the different types of weighting we have discussed, it may be helpful to glance at a few concrete cases. The object of the following experiments was to examine instances where there happened to be an objective check on the calculated ' factor-measurements ' (the weighted sums). In this way it is not difficult to show that, at any rate so far as the dominant factor is concerned, (i) equal weighting will commonly provide moderately good estimates, (ii) much better estimates can be obtained by a factor analysis, and that (iii) weighted summation with reduced self-correlations will yield better estimates than any other procedure—better even than the classical method of calculating the line of closest fit (i.e., weighted summation with unity in the diagonal).

An Example with an External Criterion. Five children, aged 12 to 15, were given a set of 80 boxes of the same size and shape, but different weight. The weights varied from 150 to 250 grains (about 10 to 16 grams). Each child was asked to assess the weight of the boxes ; and the investigator's problem was to ascertain how closely the estimates derived by averaging or otherwise combining the children's judgments would approximate to the true values, when the mode of combination was determined without reference to those values.

TABLE I. INTER-CORRELATIONS BETWEEN JUDGMENTS OF WEIGHTS

Persons	L. T. R.	F. B.	A. M.	J. C. H.	W. R.
L. T. R. ..	(·931)	·636	·531	·407	·238
F. B.	·636	(·948)	·575	·354	·152
A. M. ..	·531	·575	(·872)	·178	·131
J. C. H. ..	·407	·354	·178	(·706)	·053
W. R. ..	·238	·152	·131	·053	(·553)

[1] The last two types of unit weighting are of special interest. The first is the method of weighting implicitly adopted in analysis of variance : (see Burt, 16, p. 7). The second is a simplified form of the old triangular reduction discussed by Lagrange and Hermite. It is particularly useful as a preliminary reduction to simplify subsequent calculations. It also has a practical application in cases where (after suitable rearrangement) later variables are dependent on earlier but earlier are not dependent on later : (see Burt, 10, p. 307 and refs.). Its introduction into correlation work is due indirectly to Karl Pearson (cf. *Biometrika*, XV, p. 135, and XXIV, p. 422).

As with Galton's test-weights, the successive increments were geometrical [1]; and the set was so selected that, judged by their logarithmic values, the frequency-distribution of the weights would approximate to a normal curve. They were presented in a randomized order. Before lifting the first, each subject was given three standard boxes which he was told weighed 100, 200, and 300 grains. To allow for Weber's law both true values and assessments were converted to terms of a logarithmic scale, and were then reduced to standard measure.[2]

Owing partly to the comparatively small differences in the weights and partly to differences in care or discriminative capacity among the individual children, the agreement between their judgments is by no means close. This is shown by the inter-correlations,[3] which are set out in Table I, and give what may be called the 'inter-personal reliability' of the assessments. With 80 items the standard error of a zero correlation is approximately $\pm$ 0·112, and any coefficient over 0·217 may be accepted as significant. To obtain a measure of personal error or 'intra-personal reliability' for each child, the experiment was repeated with each subject after an interval of a week; and a correlation calculated between his judgments as given at two successive trials—a procedure originally suggested by Clark Wissler.[4] To check the weighting equations derived from the first experiment the same children repeated the test with a second set of boxes ranging from 200 to 300 grains.

Agreement of Individual Assessments with the Criterion. (a) *External Criterion.* Each child's assessments were first correlated with the 'true weights.' The coefficients are shown in the first column of Table IV. It will be seen that the correlations range from 0·196 (which is scarcely significant) to 0·841. Two out of the five children not only have reliabilities well over 0·90, but also appear able to arrange the boxes very satisfactorily in order of weight. Indeed, quite a good estimate would be obtained by accepting the judgments given by L. T. R. alone.

It would, however, seem reasonable to anticipate a more accurate estimation, if we relied, not on the judgments of a single individual, however good, but on the composite 'index' (as Pearson would have called it), which can be obtained by averaging the judgments of two or more individuals. Indeed, it is commonly supposed that the larger the number of assessments available for averaging, the greater is the accuracy of the average so obtained. First of all, therefore, let us try averaging the judgments obtained from all the judges. On calculating a straight average for each of the 80 boxes, we obtain a correlation with the true values of only 0·844. (The same coefficient is reached if we use equation (vi) above.) This figure exceeds that obtained with the best of the five judges by an insignificant fraction only; and is not nearly as high as that which would be obtained by averaging the assessments of the best two judges (0·902). It is clear that, by averaging all the assessments indiscriminately, we have allowed the inaccuracies of the poorer subjects to counterbalance the gain to be expected by pooling the judgments of the best. However, in investigations where we did not know the true values, we could not at this stage say who really were the best judges.

How, then, in the absence of any external check, can we determine their accuracy? Apparently our only means of doing so is to take the joint report of all the judges as a provisional indication, and then measure the accuracy of each one by his agreement with this internally derived criterion. We could then improve our provisional indication by substituting a weighted average on the basis of this result; and we could continue by successive approximation until we reached a stable set of figures. Factor analysis is merely a short cut to this final result.

[1] Cf. Galton, F., *Inquiries into Human Faculty*, 1883, Appendix, pp. 248–51.

[2] The experiment was suggested by the test of 'five weights' introduced into the original Binet-Simon Scale (cf. Burt, C., *Mental and Scholastic Tests*, 1921, p. 51, and 'Correlations between Persons,' *Brit. J. Psych.*, XXVIII, 1937, pp. 63-4). In practical laboratory courses for students, both at Liverpool and London, I have regularly used it as a stock exercise (usually with 10 weights only) to illustrate various correlational methods, and, in particular, the analogies and the differences in the procedure required (*a*) when we possess an external criterion and (*b*) when we have to rely only on an internal criterion. Incidentally it exemplifies the application of factor analysis to correlations between persons as well as between tests or traits.

[3] The original calculations were carried out by Mr. R. F. Chalmers, whom I have to thank for permission to use his figures; and most of the work has been checked by Mr. Summerfield, to whom I am also much indebted. I should add that more than five children took the test; but the correlation table was limited to that number to simplify both the exposition and the work.

[4] 'The Correlation of Mental and Physical Tests,' *Psych. Rev. Mon. Supp.*, III, 1901, p. 60. In all our earlier experiments with mental tests it was part of the regular routine to repeat the tests, and calculate the reliability coefficients. In theory a more satisfactory derivation of the formulæ for factor analysis becomes possible, when values for the errors are assumed to be known; but in practice it is doubtful whether the estimates would thereby be appreciably improved, particularly when only one factor has to be determined (cf. 3, pp. 112, 157).

(b) *Internal Criterion.* Let us now imagine that we present a list of figures for the five sets of standardized assessments to an investigator who knows nothing whatever about their actual nature or source (he might, for example, suppose they were marks for 80 pupils in five subjects of the elementary curriculum) ; and let us ask him to determine measurements for as many independent factors as will reasonably account for the variation in the figures.

It is evident that he could not in any case determine measurements for *more* than five common factors. If he follows Pearson's procedure, he will take the variances or self-correlations to be unity, calculate his factor-saturations by weighted summation, and so obtain weighting equations for the maximum number of components, namely, five : the matrix of regressions or weights for determining the measurements required will then be simply the inverse of the factor-matrix. The factor-saturations thus obtained are shown in Table II A. It will be seen that the first component accounts for about 49 per cent. of the total variance, and that the other components account for 19, 17, 8, and 7 per cent. respectively.

TABLE II. SATURATION COEFFICIENTS

Persons	A. For Observed Table					B. For Smoothed Table				
	I	II	III	IV	V	I	II	III	IV	V
L. T. R. ..	·861	·005	·018	−·307	−·410	·865	−·062	−·138	−·268	−·398
F. B.	·849	−·096	−·156	−·255	·428	·835	−·070	−·156	−·333	·402
A. M.	·754	−·020	−·481	·440	−·066	·747	−·097	−·304	·585	·032
J. C. H. ..	·556	−·354	·717	·221	·036	·550	−·201	·803	·106	·011
W. R.	·321	·907	·252	·073	·062	·316	·940	·118	·040	·007
Variance ..	2·443	·957	·834	·406	·360	2·405	·942	·794	·538	·321

TABLE III. CHI-SQUARED ANALYSIS [1]

(i) k	(ii) Latent Root (λ)	(iii) R_q	(iv) $\log_e R$	(v) χ^2 ($74{\cdot}5 \log_e R_q$)	(vi) D.F.	(vii) Borderline	(viii) χ^2 (Diffs.)	(ix) D.F.	(x) Borderline
0	2·443	·2870	−1·2483	93·0	10	18·3	66·8	4	9·5
1	·957	·7037	−0·3514	26·2	6	12·6	10·2	3	7·8
2	·834	·8072	−0·2142	16·0	3	7·8	15·8	2	6·0
3	·406	·9974	−0·0026	0·2	1	3·8	0·2	1	3·8
4	·360	1·0000	−0·0000	0·0	0	—	—	—	—
Total	5·000	—	—	—	—	—	93·0	10	—

[1] *Note to Table III.* In col. (i), k denotes the number of factors that have been extracted. Write $q = p - k$ for the number of roots still to be extracted. Then in col. (iii),

$$R_q = q^q \frac{\lambda_{k+1}\lambda_{k+2}\ldots\lambda_p}{(\lambda_{k+1}+\lambda_{k+2}+\ldots+\lambda_p)q} = q\text{th power of } \frac{\text{Geom. mean of last } q \text{ roots}}{\text{Arith. mean of last } q \text{ roots}},$$

col. (v) gives chi squared for the residual correlations remaining after k factors have been removed ; the multiplier $= N - p - \frac{1}{2} = 80 - 5 - \frac{1}{2} = 74{\cdot}5$, where N = the number of items over which the correlation extends, and p the number of persons correlated ; col. (vi) gives the corresponding degrees of freedom, and col. (vii) the borderline values of chi squared at the 5 per cent. level ; col. (viii) gives chi squared for the separate factors (obtained by taking successive differences from col. (v)) ; and cols. (ix) and (x) give the corresponding degrees of freedom and borderline values.

I am indebted to Professor Bartlett for an advance note giving his formula for testing significance. His method has been described more fully in his paper published in this issue (see pp. 77–85). My own method of expressing his formula was intended partly to suggest a routine method of computation when all the latent roots are known, and partly to indicate certain analogies with other statistical tests (e.g., for homogeneity of variance). Note that, when the side correlations are approximately hierarchical, $|R|$ can be calculated by the formula $|R| = (1 + f'U^2f) \cdot |U^2|$, where f = the vector of saturations, and U^2 the diagonal matrix of specific factor variances.

Now with a ' principal axes ' analysis, based on the total variance (unity in the diagonal of the correlation), we can apply the chi-squared test in the form suggested by Bartlett (Table III). From the figures it will be seen that three factors at most would be sufficient to account for all that is significant in the correlation structure. Further, on examining the pattern [1] presented by the factor-saturations, it becomes clear that the four bipolar factors can represent little else than the results of forcing specific factors into the form of common factors. This can best be shown if we reconstruct a smoothed correlation matrix from the saturations for the first factor alone, and then factorize it by Pearson's method with unity in the diagonal : the results are given for comparison in Table II B, where the artificial nature of the resulting factor pattern is very plain. In Table II A there is certainly a suggestion of a small but genuine bipolar factor contrasting F. B. and A. M. with the other three subjects ; but, until this has been disentangled from the effects of the specific factors, its significance must remain highly doubtful. Our investigator can therefore, with little or no hesitation, assume the existence of a dominant factor with positive saturations throughout, but that is all ; and the hypothesis which at this stage appears to be at once the most economical and the most plausible will be that the five sets of correlated assessments are explicable as the result of (i) a single general factor and (ii) five specifics.[2] Only if this factor-structure is found unavoidably to leave residuals that are definitely significant shall we be warranted in postulating (iii) one or more supplementary factors in addition.

His next step, therefore, will be to test his hypothesis by finding how closely he can fit his data with saturations deduced from it. To find the saturations yielding the closest fit he will again use the method of weighted summation, but will now insert appropriately reduced self-correlations into the diagonal of the correlation matrix.

The principle of economy demands that he should begin by basing his estimated self-correlations on the assumption of a single common factor first of all, and only introduce further factors when the evidence obliges him to ' multiply the entities ' he is postulating. Table I can be fitted reasonably well with one common factor alone. The largest residual is −0·088 ; and this can have no statistical significance. Had the first set of residuals been statistically significant (as it might easily have been with a larger group), then he would have to augment his self-correlations accordingly, and commence his calculations afresh. In cases like the present, where probable errors are fairly large, he might in actual practice be content with simple summation instead of weighted. On the other hand, if he wished to adopt a more rigorous treatment, he would first ' correct for unreliability ' (thus taking a matrix of maximum reliability) ; or, if the reliabilities were unknown (as is more usually the case), ' correct for specificity ' (16, p. 113 ; the formulæ are the same, except that in the former case r_{ii} denotes reliabilities, in the latter reduced self-correlations).

Our investigator's ultimate aim, however, is to estimate the appropriate measurements for the first (and only significant) factor. This factor provides us with a criterion based on internal evidence ; and, so far as it can be accepted as a trustworthy substitute for the external evidence of the balance, we may take the ' saturation coefficients ' for this factor as estimates of the probable correlation of each judge's assessments with the ' true weights.'

Accordingly, we have next to consider the problem of finding the most suitable weights for this purpose. Let us compare the results furnished by the various methods available for calculating (*a*) these estimated correlations between the assessments and the criterion, and (*b*) the regression coefficients to which these correlations would severally lead.

Judges' Correlations with the Criterion. The saturation coefficients are shown in the last five columns of Table IV.

[1] Cf. 16, p. 112, for a fuller explanation of this point. I have to acknowledge the assistance of Mr. Summerfield, who has been good enough to check my calculations.

[2] Had he been allowed some antecedent knowledge of the source, this is the factor-pattern that I imagine he would regard as most probable on *a priori* grounds. If the data had been procured as a result of some hypothesis that he himself had set up (the ideal procedure in a factorial study), that hypothesis, I imagine, would have begun by postulating a three-factor scheme—with basic, group, and specific factors ; his task would then have been to verify or disprove this hypothesis either wholly or in part, and, if it were wholly verified, to measure the relative importance of the several factors by calculating the most likely estimates for their variances. The bipolar or group factors would, of course, represent the contrast between different types of judge—e.g., those who are more or less subject to weight-illusions, or possibly Binet's *interprétateurs* and *simplistes ;* and the judges might even have been deliberately chosen in the hope of eliciting such type-differences if they existed (for the existence of such factors, see *Brit. J. Educ. Psych.*, XIX, 1949, p. 110, and refs.).

Five methods of analysis have been employed. (i) Weighted summation, with self-correlations of unity (*W.S.u.s.*) ; this, it will be remembered, adopts as the internal criterion Pearson's line of closest fit. (ii) Simple summation, with self-correlations of unity (*S.S.u.s.*) ; here the criterion is the straight unweighted average of the standardized assessments. (iii) Weighted summation with correction for specificity (*W.S.c.s.*). (iv) Weighted summation with reduced self-correlations, but without correction for specificity (*W.S.r.s.*). (v) Simple summation with reduced self-correlations (*S.S.r.s.*).

TABLE IV. AGREEMENT OF INDIVIDUAL ASSESSMENTS WITH CRITERION

Persons	(*a*) With External Criterion	(*b*) With Internal Criterion (Saturation Coefficients)				
		(i) *W.S.u.s.*	(ii) *S.S.u.s.*	(iii) *W.S.c.s.*	(iv) *W.S.r.s.*	(v) *S.S.r.s.*
L. T. R. ..	·841	·861	·829	·818	·838	·863
F. B.	·790	·851	·801	·805	·810	·815
A. M. ..	·634	·753	·712	·669	·642	·625
J. C. H. ..	·402	·556	·587	·427	·414	·399
W. R. ..	·196	·320	·464	·238	·221	·214
Discrepancy (Root-mean-square)	·000	·107	·149	·030	·015	·018

TABLE V. WEIGHTING COEFFICIENTS (PARTIAL REGRESSIONS)

Persons	(*a*) Based on External Criterion	(*b*) Based on Internal Criterion				
		(i) *W.S.u.s.*	(ii) *S.S.u.s.*	(iii) *W.S.c.s.*	(iv) *W.S.r.s.*	(v) *S.S.r.s.*
L. T. R. ..	·534	·353	·295	·421	·480	·532
F. B.	·346	·350	·295	·378	·400	·403
A. M. ..	·154	·308	·295	·204	·141	·101
J. C. H. ..	·040	·228	·295	·081	·048	·022
W. R. ..	−·002	·130	·295	·048	·029	·011
Correlation						
(i) Main Set	·914	·885	·844	·907	·910	·908
(ii) Second Set	·862	·781	·837	·885	·902	·873

Let us first enquire how far the hypothetical correlations with the internal criterion afford a trustworthy indication of the judges' observed correlations with the external criterion, i.e., the true weights. The closeness of the approximation can best be assessed by computing the root-mean-square of the several discrepancies (see last line of Table IV). It is plain that the nearest figures are the saturations furnished by the method of weighted summation with reduced self-correlations. Those found by taking self-correlations of unity are decidedly inaccurate : the effect of this procedure is to give a deceptively high figure for the accuracy of the poorest judges (especially J. C. H. and W. R.). The worst figures of all are those furnished by simple summation with self-correlations of unity.

Estimation of the Criterion by a Weighted Combination of the Judges' Assessments. As we have seen, neither the assessments of the most accurate judge nor yet a simple average of all the assessments can be expected to yield the best available estimates for the weights of the several boxes. Our problem now is to enquire how far such estimates are improved by the different methods of weighting that are available. For this purpose the necessary weights or loadings will be calculated from the correlations or saturations by the usual formula for partial regressions.

If we already knew the true values for each box, we could calculate the regressions from the observed correlations : the equation to use would be $w' = r'R^{-1}$, where r stands for the first column in Table IV and R the correlation matrix in Table I with unity in the diagonal. The loadings so

obtained are set out in the first column of Table V. The figures indicate that A. M.'s judgments deserve less than half the weight of F. B.'s and little more than a quarter of that of L. T. R.'s, while the assessments of J. C. H. and W. R. have no significance at all. In general, of course, the regressions obtained in this way from a direct comparison with the true values will furnish the highest multiple correlation. Here, it will be seen, the multiple correlation amounts to 0·918.

Now, as I ventured to suggest in my earlier papers (4, cf. 10 and refs.), even where we have no external criterion we can still apply the principles of multiple correlation to determine measurements for our so-called 'factors.' The regressions will be calculated by a similar formula ($w' = f'R^{-1}$, where f' will now denote any of the last five columns of Table IV). And generally, to determine the relations between the internal criteria (P) and an external criterion (c), we may use the same weights as those I have elsewhere suggested for computing the correlations between a set of factors and *any* external variable: viz., (in the notation of my previous papers) $W = F^{-1}$ (if F is non-singular), or $F'R^{-1}$ (if it is not), i.e., $V^{-1}F'$ (if weighted summation has been used), or $v^{-1}f'$ (if only one factor is required). To transform F to r_{mc}, we adopt the usual type of post-multiplier, viz., $F^{-1}r_{mc} = F^{-1}Mc' = Pc' = r_{pc}$, making the same substitutions as before if F is non-singular (cf. 10 and Appendix, p. 124). The weighting coefficients thus obtained are shown in the corresponding columns of Table V. The correlations with the *true* values, calculated by using these weights, are appended in the last line but one. Confirmatory validation, as usual, has been sought by trying the same weighting equations with a *second* sample: see last line of table.

It is clear that simple averaging (*b* ii) yields the least satisfactory result. As we have seen, it gives equal weights to all judges. Even the Pearsonian method for finding the 'line of closest fit' (*b* i) tends in the same direction. On the other hand, the methods of simple and weighted summation *with reduced self-correlations* (*b* iii, iv, and v) here supply almost as close an agreement with the true values as was obtained from the partial regressions based on the true values themselves. In the present case the differences between the last three methods are too slight for any generalization to be deduced from the figures alone.

Supplementary Experiments. To show that these results are in no way exceptional, but hold good in widely different fields, I append the figures calculated by similar methods from six other enquiries.

The tasks were (1) to judge the weights of boxes (an experiment identical with the above but carried out with adults); (2) to judge the length of lines; (3) to judge the areas of rectangular figures; (4) to judge the length of time-intervals; (5) to judge the number of spots on cards shown tachistoscopically; (6) to guess the heights of familiar buildings or monuments. In the first and last of these experiments adults (students) acted as judges; in the other three, children aged 12 to 14. Table VI shows the final correlations between the 'true values' and the estimates given by various linear combinations of the assessments made by all the judges.[1]

TABLE VI. CORRELATIONS BETWEEN JUDGMENTS AND TRUE VALUES

Test Material	No. of Items	No. of Judges	(*a*) Based on External Criterion	(*b*) Based on Internal Criterion (i) *W.S.u.s.*	(ii) *S.S.u.s.*	(iii) *W.S.c.s.*	(iv) *W.S.r.s.*	(v) *S.S.r.s.*
Weights ..	30	10	·963	·954	·937	·966	·958	·972
Lines ..	50	6	·927	·813	·748	·913	·930	·895
Areas ..	20	8	·668	·623	·605	·545	·584	·491
Time-intervals	50	6	·751	·680	·663	·738	·742	·731
Spots ..	100	7	·842	·776	·736	·791	·833	·802
Buildings..	20	8	·765	·719	·721	·754	·772	·768
Average ..			·819	·761	·735	·785	·803	·777

[1] The experiments with children were carried out and analysed by Mr. R. F. Chalmers, Mr. B. C. Murray, and Miss K. Willis, to whom I am indebted for the figures here given. The number of judges is in every case rather small: that is almost inevitable, owing to the labour involved in calculating an inverse for larger correlation tables.

The results may be summarized as follows. Where the number of items is small, or where the number of judges is comparatively large, and particularly where all the judges have much the same degree of efficiency, there is little to choose between the various methods of weighting, and not much is lost if no weighting whatever is employed. But where the number of items is large enough to reduce the probable error to small dimensions, and particularly where the number of assessments to be averaged is small, there the use of an accurate method of weighting is plainly of special importance. In such cases the best procedure would seem to be the method of least squares (weighted summation) with reduced self-correlations. The method of simple summation usually yields very similar results. There is one exception—the experiment with areas. Here (as the introspections show) at least one large supplementary factor is almost certainly operative (representing subjective differences in mode of assessment). With so few cases the residuals for such a factor were not statistically significant, and tended to be somewhat erratic. However, if a second factor were allowed for in the self-correlations, these would approach nearer to unity. Hence the apparent exception is in keeping with the general rule. It may seem surprising that the correction for specificity does not (except in one dubious case) improve the resulting figures. That is doubtless due to the fact that the external criterion is not absolutely identical with the internal.

Is Weighting Really Necessary? In discussing the weighting of test items for validity, Guilford apparently concludes that " what few results we have point definitely to the inference that such weighting is not worth the trouble " (*loc. cit.*, p. 448). Can we then lay down any criterion for deciding when, if ever, such calculations are really worth while? I would suggest, as a sound practical rule, that differential weighting may be regarded as unnecessary when the correlation between the weighted and unweighted assessments is significantly larger than the reliability coefficient for the weighted assessments.

Thus, to take the first experiment described above, we find that, with method (*b* iv) (weighted summation) the reliability of the composite figures would be 0·969 ; the correlation between these composite figures and those given by the unweighted average of the standardized judgments is 0·941. In this case, therefore, the use of differential weighting would certainly seem desirable. Whenever the number of tests to be averaged is small, and more especially when their inter-correlations are low, the problem of weighting ought in my view never to be dismissed without at least a rough preliminary investigation of the point.

V. SUMMARY AND CONCLUSIONS

1. An attempt has been made to investigate the conditions affecting the results obtained with different methods of weighting. The problem leads to a consideration of (*a*) *a priori* weighting as the oldest form, (*b*) random weighting as the most general form, (*c*) equal weighting as the simplest form, and (*d*) differential weighting as determined empirically by various factorial procedures.

2. *A priori* or subjective weighting may be necessary (i) where questions of value are concerned, or (ii) where the criterion to be predicted is genuinely composite. Where the criterion is factual and essentially unitary (though assessed by multiple observations), *a priori* weighting is generally worthless.

3. Equations have been derived to express the correlations obtained both by random weighting (such as tends to be produced by accidental differences in standard deviations) and by equalized weighting (where such accidental differences have been eliminated). The formulæ indicate that the validity coefficients so obtained increase (i) with the number of weighted traits and (ii) with the size of the inter-correlations between them. It is concluded that the choice of the best possible set of differential weights is of special importance when the traits are few in number and differ widely in their diagnostic value.

4. A series of illustrative experiments is described in which an external as well as an internal criterion was available. The results indicate that, when an external criterion is not available, the best weights will be obtained with a factor analysis carried out, not by the classical method giving the ' lines of closest fit ' (i.e., the method of principal components with self-correlations of unity), but by weighted summation with reduced self-correlations. Where the correlation matrix does not depart

appreciably from a rank of one, the further corrections for specificity or reliability introduce no discernible improvement, and simple summation gives almost as good a result as weighted summation.

5. More generally it has been shown (i) that *any* method of weighting will yield a ' factor ' in the accepted sense of the term, and that the extraction of such a factor always reduces the rank of the matrix of test-measurements and the matrix of the correlations (or the residual matrices) by one ; (ii) that the choice of a weighting system which yields *orthogonal* factors leads to particularly simple formulæ and convenient working procedures. One such system is obtained if the weights selected are numerically equal, but differ in sign-patterns for successive factors, i.e., if the standardized measurements are merely summed algebraically as they stand. This is the weighting adopted in analysis of variance. When the number of traits to be summed is small, this type of weighting does not yield altogether satisfactory results ; but when the number is large, it yields a good approximation to the best result, particularly for the first or ' general ' factor.

6. The weighting coefficients deduced from a particular sample necessarily give better predictions for that sample than for any other, especially when the number of predictors is large. Hence an empirical weighting equation should always be checked by application to a second experimental group. When this is done, weighted summation with reduced self-correlations again seems to furnish the most stable results.

7. A criterion has been suggested, based on the reliability of the composite result, to indicate when differential weighting should be substituted for equal weighting.

REFERENCES

1. Edgeworth, F. Y. (1888). ' The statistics of examinations,' *J. Roy. Stat. Soc.*, LI, 599–635.
2. Johnson, W. W. (1892). *The Theory of Errors and the Method of Least Squares.* Oxford : University Press.
3. Burt, C. (1909). ' Experimental tests of general intelligence,' *Brit. J. Psych.*, III, 94–177.
4. Burt, C. (1917). *Distribution and Relations of Educational Abilities.* London : King.
5. Bowley, A. L. (1920). *The Elements of Statistics.* London : King.
6. Burt, C. (1921). *Mental and Scholastic Tests.* London : King.
7. Keynes, J. M. (1921). *A Treatise on Probability.* London : Macmillan.
8. Edgeworth, F. Y. (1925). ' Index numbers.' *ap.* Palgrave, *Dictionary of Political Economy.*
9. Brunt, D. (1931). *The Combination of Observations.* Cambridge : University Press.
10. Burt, C. (1935). ' Memorandum on Methods of Factor Analysis ' ; subsequently published in Hartog, P., and Rhodes, E. C. *The Marks of Examiners*, pp. 245–314. London : Macmillan.
11. Monroe, W. S., and Engelhart, M. D. (1936). *The Scientific Study of Educational Problems.* London : Macmillan.
12. Frazer, R. A., Duncan, W. J., and Collar, A. R. (1938). *Elementary Matrices.* Cambridge : University Press.
13. Kendall, M. G. (1939). ' The geographical distribution of crop productivity in England,' *J. Roy. Stat. Soc.*, CII, 21–48.
14. McClelland, W. (1942). *Selection for Secondary Education.* London : University of London Press.
15. Burt, C. (1948). ' A comparison of factor analysis and analysis of variance,' *Brit. J. Psych., Stat. Sect.*, I, 3–26.
16. Burt, C. (1949). ' Alternative methods of factor analysis, and their relation to Pearson's method of principal axes,' *Brit. J. Psych., Stat. Sect.*, II, 98–121.
17. Kendall, M. G. (1950). ' Factor analysis as a statistical technique,' *J. Roy. Stat. Soc.* (in the press).

APPENDIX

1. In the text (p. 108) I have contended that the real or operative weight of any component is to be measured (as in factor analysis) by its contribution to the total variance. Thus, if

$$y = B_1X_1 + \ldots + B_nX_n = B_1\sigma_1x_{1i} + \ldots + B_n\sigma_2x_{ni} = U_1x_1 + \ldots + U_nx_n ,$$

where B_j = the *assigned* weight and $U_j = B_j\sigma_j$ = the *actual* weight, then

$$\begin{aligned}\sigma_y^2 &= B_1\sigma_1\textstyle\sum B_i\sigma_i r_{1i} + \ldots + B_n\sigma_n\sum B_i\sigma_i r_{ni} \\ &= U_1\textstyle\sum U_i r_{1i} + \ldots + U_n\sum U_i r_{ni} :\end{aligned}$$

so that the contribution of j depends on (i) B_j, (ii) σ_j, and (iii) the r_{ij}'s. If the components are in standard measurement and uncorrelated, then $\sigma_y^2 = U_1^2 + \ldots + U_n^2$, i.e., the contribution of each component varies with the square of its weight. If the components are correlated, and given equal weights of unity, then $\sigma_y^2 = \sum r_{1j} + \ldots + \sum r_{nj} = r_{1y} + \ldots + r_{ny}$, i.e., the contribution of each component varies with its correlation (or 'saturation') with the composite measurement, y. But with differential weighting the saturations will not be proportional to the regressions; and sometimes, as with 'suppressor' variables, a test with a positive saturation may require a negative weight.

2. Where the final criterion is external to the battery, the best method of weighting is that described in the theory of multiple correlation. Where no such criterion is available, then (as stated above) the same kind of proof leads to the formulæ for factor analysis by weighted summation. But, it may be asked, when *both* are available, how are we to determine the relation between the two?

There are two ways in which the relation between a factor and an external criterion may be assessed. We may correlate either (i) the factor-*measurements* (p) with the criterion *measurements* (c), or (ii) the factor-*saturations* (f') with the correlations between the criterion and the several assessments (r_{mc}). The former is in my view the proper procedure; the latter provides a convenient short cut, but has sometimes been adopted without a real understanding of its implications.[1]

(i) If p is in unitary standard measure, then the correlation between the measurements will be $r_{pc} = pc' = w'Mc' = v^{-1}f'r_{mc} = v^{-1}f'r_{mc} = (f'f)^{-1}f'r_{mc}$. But if p consists of *estimated* measurements, then its variance will be $w'f = f'R^{-1}f$; and this is only unity when f has been derived by factorizing the correlation table with self-correlations of unity. Hence the formula used in the text is $w'r_{mc} \div \sqrt{w'f}$.

(ii) The 'unadjusted correlation' between the saturations and the criterion correlations (r_{mc}) will be $\dfrac{f'r_{mc}}{\sqrt{(f'f)}\sqrt{(r_{mc}'r_{mc})}} = v^{-1}f'r_{mc}$, if we take $r_{mc}'r_{mc}$ as $= f'f$ approx. It will be seen that, with this assumption, the two methods of calculation lead to the same result. Moreover, when the 'unadjusted correlation' = 1, then the saturations and criterion correlations are proportional. If, however, we desire to compare *several* sets of saturations (e.g., sets obtained by different procedures), then it would seem better to use the square-sum of the differences, though this is again closely related to the unadjusted correlation.

3. The full proof of the equations (iii) and (iv) given on p. 111 was given in my original (unpublished) memorandum (cf. 10, p. 246): though lengthy, it is straightforward. The reader will find no difficulty in reconstructing the argument for himself. The essential steps may be summarized as follows.

Using the notation suggested on p. 111 consider the product-moment correlation between the weighted sums, y and z, defined by equations (i) and (ii); write the correlation $r_{yz} = \dfrac{N}{D_u D_w}$, where N and D_uD_w denote the numerator and denominator respectively.

Then, since the measurements are in unitary standard measure,

$$N = \sum_i^n U_iW_i + \sum_{i \neq j}^{n(n-1)} U_iW_jr_{ij} .$$

Putting $U_i = \overline{U} + u_i$ and $W_i = \overline{W} + w_i$, we have

$$\sum_i^n U_i W_i = n\,\overline{U}\,\overline{W} + \sum u_iw_i .$$

Now U_i and W_i, having been chosen at random, must have only non-significant correlations with each other and with the values they are used to weight. Further, unless there are wide divergences among the inter-correlations, we may take $\sum U_i W_j r_{ij} = \bar{r} \sum U_i W_j$ approximately (where $\bar{r}$ denotes the average value of r_{ij}, $i \neq j$). Hence $\sum_{i \neq j}^{n(n-1)} U_iW_jr_{ij} = n(n-1)\,\overline{U}\,\overline{W}\bar{r} + \bar{r}\sum u_i w_u$ approximately.

[1] For the formulæ for calculating the correlation between factors and external variables, see 10: also *J. Psychol.*, VI, 1938, p. 355, eq. (vii)–(ix).

But, for the same reason as before, $\sum u_i w_j$ must tend to zero. Hence

$$N = n\,\overline{U}\,\overline{W}\,(1 + n - 1\,\bar{r}) \text{ approximately.}$$

Similarly $D_u = n\{\overline{U}^2\,(1 + (n-1)\,\bar{r}) + (1 - \bar{r})\,\sigma_u^2\}$, with a corresponding expression for D_w.

Now, the correlation r_{yz} depends solely on the *relative* proportions of the weights; hence it will be unaffected if we multiply the first set of weights by some constant, so that $\overline{U} = \overline{W}$. Then, since by hypothesis $\sigma_y^2 = \sigma_z^2 = 1$, we have $\sigma_u^2 = \sigma_w^2$; and therefore

$$r_{yz} = \frac{n\overline{W}^2\,(1 + n - 1\,\bar{r})}{n\{\overline{W}^2\,(1 + n - 1\,\bar{r}) + (1 - \bar{r})\,\sigma_w^2\}}$$

$$= \frac{n\,\bar{r} + (1 - \bar{r})}{n\,\bar{r} + (1 - \bar{r}) + (1 - \bar{r})\,\dfrac{\sigma_w^2}{\overline{W}^2}}.$$

Dividing out we obtain

$$r_{yz} = 1 - \frac{(1 - \bar{r})}{n\,\bar{r}} \cdot \frac{\sigma_w^2}{\overline{W}^2} + \text{terms in } \left(\frac{1}{n}\right)^2 \text{ and higher powers.}$$

If we prefer to retain the original values for the weights, so that $\overline{U} \neq \overline{W}$, and $\sigma_u \neq \sigma_w$, then $D_u D_w$ can be factored by 'completing the square'; and we reach the same general expressions as before except that for $\dfrac{\sigma_w^2}{\overline{W}^2}$ we substitute $\frac{1}{2}\left\{\dfrac{\sigma_u^2}{\overline{U}^2} + \dfrac{\sigma_w^2}{\overline{W}^2}\right\}$.

If we suppose that the inter-correlations are too widely divergent to substitute $\bar{r}\sum U_i W_j$ for $\sum U_i W_j r_{ij}$, the algebra becomes somewhat heavier; but it is not difficult to show that the effect is merely to diminish the value of the negative term, so that it tends towards $\dfrac{1}{n\,\bar{r}} \cdot \frac{1}{2}\left\{\dfrac{\sigma_u^2}{\overline{U}^2} + \dfrac{\sigma_w^2}{\overline{W}^2}\right\}$.

4. The other equations given in the text are readily reached by the use of matrix notation. The following results may be noted. If y and z represent factor-measurements, and the vectors of weights, w_y and w_z (say), are based on factor-saturations, f_y and f_z, then

$$r_{yz} = \frac{w_y'\,R\,w_z}{\sqrt{w_y'\,R\,w_y \;.\; w_z'\,R\,w_z}} = \frac{f_y'\,R^{-1}\,f_z}{\sqrt{f_y'\,R^{-1}\,f_y \;.\; f_z'\,R^{-1}\,f_z}}.$$

If w_y has been obtained by a Pearsonian analysis, $r_{yz} = f_y'\,w_z$, assuming w_z is standardized.

If, however, $w_z = [1, 1, \ldots 1] = w_0$ (in the usual notation), then $r_{yz} = \dfrac{f_y'\,w_0}{\sqrt{w_0'\,R\,w_0}}$.

NOTE ON SIR CYRIL BURT'S PAPER ON DIFFERENTIAL WEIGHTING

By GODFREY THOMSON
Moray House, University of Edinburgh

I fear the preceding paper may be misleading unless it is read with the following warning in mind. I am thinking more particularly of the experiments with lifted weights, etc., in which it was found that the method of weighted summation *with reduced self-correlations* provides almost as good an agreement with the true weights as would be obtained from the partial regressions based on those true weights ; and the method of simple summation provides estimates that are almost as good (pp. 121, 123). The conclusion implicitly drawn by the writer is that ' reduced self-correlations,' which most of us call ' communalities,' are justifiably used in analyses of psychological tests—a conclusion about which I am very sceptical. In the experiment with lifted weights we *know* that the correlations are entirely due to one factor, the weight, which is the only thing common to the five persons making the judgments. In that case, it seems to me obvious that it is proper to reduce the rank of the matrix, by the insertion of communalities, until it is of rank one, or as nearly rank one as the inevitable blemishes will permit. Burt's experimental confirmation of this is interesting, but the result seems to me to be a foregone conclusion. To reverse this argument, however, seems to me to be dangerous. In a psychological set of correlations between tests (and the persons in Burt's experiments correspond to tests in the usual psychological analysis, while the weights in his experiment correspond to the children in the usual psychological experiment), even when the rank can be reduced by communalities to one, we do not know that there is only one factor causing the correlations. The correlations may have been caused in many other ways, and in particular might have been caused by tests which call upon random samples of a large number of qualities whether inherited or acquired (genes, bonds, pieces of information, fragments of skill, what you will). It is true that when the rank can be reduced to one, the tests have acted " as if " they had only one factor in common, together with a large specific factor in each test, but this is not necessarily the case, and it seems to me that to insert the minimum possible communalities in the diagonal cells is to make the assumption at once that there is only one common factor and that each test has a very large specific factor, and to make this assumption solely on the mathematical grounds that the low rank can be attained, or very nearly attained, without any psychological consideration whether these large, these enormously large, specific factors actually exist.

It seems to me that we are only justified in using minimum communalities if (*a*) we know that the correlations are only due to a certain small number of common causes, as was the case in Burt's experiment, or (*b*) we think it worth while to make the assumption of a similar number of hypothetical common causes for convenience of thinking. In psychological experiments we can seldom, if ever, know that there is actually only a small number of common causes acting. It may be worth while to speak and to act " as if " this were the case, but what I am afraid of is that these purely hypothetical entities may be held to have a proved existence in a much more definite sense, sometimes even in a physical sense.

REFERENCES

1. Thomson, Godfrey H. (1916). *Brit. J. Psych.*, VIII, 271–81.
2. Thomson, Godfrey H. (1919). *Proc. Roy. Soc. Lond.*, A, XCV, 400–8.
3. Thomson, Godfrey H. (1927). *Brit. J. Psych.*, XVIII, 68–76.
4. Mackie, John (1928). *Brit. J. Psych.*, XIX, 65–76.
5. Mackie, John (1929). *Proc. Roy. Soc. Edin.*, XLIX, 16–37.
6. Thomson, Godfrey H. (1934). *Nature*, CXXXIV, 700.
7. Thomson, Godfrey H. (1935). *Brit. J. Psych.* XXVI, 63–92.
8. Thomson, Godfrey H. (1949). *The Advancement of Science*, VI, 23, 267–74.

A REPLY TO SIR GODFREY THOMSON'S NOTE

I am most grateful to Sir Godfrey Thomson for his comments on my paper. With his warning, and with the gist of his final conclusion, I am, I think, in general agreement. He puts his finger on what to my mind is the most vexed question at the present moment, namely, are psychologists justified in substituting 'reduced correlations' for unity in the diagonal of the matrix to be factorized? He thinks this change unjustifiable. Presumably he would revert to Pearson's procedure—though he does not explicitly say so, nor can I recollect any factorial investigation where he has himself adopted that method.

Space does not allow me to consider his arguments in detail. But I should like to make the following brief comments.

1. In the foregoing paper I was only concerned to compare methods of weighting, not to adduce evidence for the assumption of "one factor causing the correlations." I have nowhere attempted to prove such a conclusion by "reversing the argument"; and I fully agree with Thomson that it could not be done that way. If I did not insert a more emphatic caution on this point, that was simply because I had dealt with this particular criticism in the last number of this *Journal*.[1]

2. My reduction of the 'self-correlations' is based not, as Thomson seems to suppose, on considerations of rank, but on considerations of significance. (I avoided the term 'communality' just because that has become associated with the idea of reducing rank.) My general grounds for preferring reduced self-correlations to self-correlations of unity have been given in earlier papers, and were summarized on p. 113. With these grounds, however, Thomson does not deal. The object of "the experiments with lifted weights, etc.," was, as stated in the opening sentences of Section IV, merely "to illustrate the effects of different types of weighting" where an objective criterion was available, not to justify the reduction of self-correlations or communalities as a general principle. That would have been far too wide a question to re-examine in an article dealing with another problem.

3. Thomson goes on to say that "in the experiment with lifted weights we *know* that the correlations are entirely due to one factor." With this, I fear, I cannot agree. On the contrary, I strongly suspect that the correlations are due to more than one common factor. My point was, not that no further factors existed, but that, in the experiment described, the additional factors were too small to be statistically significant; and in such a case, so I should maintain, all we can take for granted is the existence of the one common factor which is established as fully significant.

4. Nor can I agree that this result was "a foregone conclusion." In fact the outcome of the experiment on judging areas seems to show that it was not. Here there *were* other factors.

5. At the same time, in thus assuming that (within the margin allowed by the probable errors) the correlations can be accounted for by means of only one common factor, I am not for one moment supposing that this single factor is therefore simple. The fact that, as Thomson suggests, a general factor like intelligence may be biologically due to a large number of genes, or neurologically due to a large number of neural bonds, does not appear to me to be in the least incompatible with its operation as a single independent factor. Sweden may be a single independent State though it consists of over six million persons.

6. Thomson would prefer to say that in cases such as I have described "the tests have acted 'as if' they had only one factor in common." With this form of expression I should heartily agree. But then, surely every empirical generalization in any and every science can state no more than this. For example, in mechanics, chemistry, and thermodynamics observable changes take place 'as if' energy was conserved; but does anyone suppose that energy and similar "purely hypothetical entities" have a "proved existence in any more definite sense"?

7. Thomson, like other critics, may have misunderstood my insistence, here and in previous papers, on the inclusion of an external criterion (such as the teacher's ranking for 'general intelligence') and on the calculation of the correlation between this criterion and the hypothetical factor-measurements. Some have apparently inferred that the object of such calculations was

[1] 'The Two-Factor Theory,' this *Journal*, II, p. 152. With the criticisms advanced in his most recent paper (8) I have already dealt in *Brit. J. Sociol.*, I, pp. 162–4. My agreement with Thomson's general view is already implied in the text, where I deprecated the common confusion between 'factors as statistical components' and 'factors as concrete entities'—"a confusion" (I added) "against which Professor Thomson has so rightly and so often protested" (p. 106): cf. also *Factors of the Mind*, pp. 213f., Section on 'The Reality of Factors.' If I am asked, how do I attempt to establish the existence of 'general ability,' I should reply, as I have so often said, that the hypothesis is based on many converging lines of concrete (non-mathematical) evidence; and I regard factor analysis merely as a way of verifying that hypothesis. Thus the argument is formally, not deductive, but *inductive*, or, as it is now the fashion to say, 'hypothetico-deductive': (cf. *Brit. J. Sociol.*, p. 162, footnote 3 and refs.). Thomson, I gather, is inclined to reject the inference: as he has said elsewhere, it is (in his view) "not necessary to postulate the existence of a general factor," and he is "not personally a believer in the existence of a faculty called 'general ability.'"

literally to identify the general factor with a ' known common cause ' (as Thomson calls it) ; and they have accordingly asked : why not go further, and *maximize* the correlation of your factor with that criterion ? My reply [1] would be that, with a properly planned experiment, the general or basic factor should be far more trustworthy than the external criterion. The calculation of the correlation between the two was intended to throw light on *both*—to correct the empirical criterion as well as to check the interpretation and labelling of the factor. Thus, in the case of ' general ability ' or ' intelligence ' (which, I take it, is the kind of ' hypothetical common cause ' Thomson has more especially in mind) I should regard the factor-measurements derived from ' a psychological set of correlations between tests ' as furnishing better estimates than the ' criterion ' provided by the teacher, though his assessments might at first sight seem more objective. May I add that in my view the ' assumption of such a hypothetical common cause ' is made, not merely ' for convenience of *thinking*,' but also for the convenience of *practical* work—e.g., allocation to schools or predictions of future performances ?

CYRIL BURT.

ERRATA

1. On p. 29 of the previous issue (Vol III, Part I) in Table IV A, col. 3 (bipolar factor III), the signs for the first 13 traits should be negative, and for the next 7 positive (the division thus corresponding with the division between group factors E and G) ; the next 9 should also be positive, and the last 9 negative (to agree with the allocation of signs adopted on p. 32).

2. On p. 67 in Table XVIII B, the figures printed in the first two rows and columns correspond with an alternative division of the first four tests (namely, into 1–2 and 3–4). To correspond with the division adopted in XVIII A and C (into 1 and 2–4), Table XVIII B should therefore read as follows :

	1	2–4	5–7	8–10
1	·7805	1·5695	1·3995	1·5105
2–4	1·5695	3·4206	2·7605	2·9795
5–7	1·3995	2·7605	3·1996	2·7200
8–10	1·5105	2·9795	2·7200	4·9601

[1] I should also answer that, as a rule, it would be better to maximize the correlation of the criterion with the *tests* than with the *factor* (i.e., to use ordinary multiple correlation) : but here I am concerned only with aims, not with methods. I may add that, when I first suggested the device of correlating internal factors with external criteria, I used *two alternative criteria :* the object was to see with which of these two abilities the general factor could best be identified. And I went on to emphasize that the factor itself, " though single is not simple, but complex " (4, 1917, p. 55).